D4.9

Specification of protocol and APIs for research health data sharing - V2

ABSTRACT

This deliverable provides specifications for the Research Data Sharing (RDS) Protocol that governs the process of collecting health data from citizens' smart health data, contained on their mobile devices, for the purposes of cross-border medical research. The deliverable defines the scope of the protocol within the entire process of setting up and preparing a research study. It also defines the actors involved, the process, as well as the underlying programming interfaces, high-level system components, and data models.

Delivery Date	9th August, 2021
Work Package	WP4
Task	T4.3
Dissemination Level	Public
Type of Deliverable	Report
Lead partner	UNITN

DRAFT

InteropEHRate project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106.

This document has been produced in the context of the InteropEHRate Project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106. All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.

This work by Parties of the InteropEHRate Consortium is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

DRAFT

CONTRIBUTORS

	Name	Partner
Contributors	Gábor Bella	UNITN
Contributors	Simone Bocca	UNITN
Contributors	Stefano Dalmiani	FTGM
Contributors	Francesco Torelli	ENG
Contributors	Marcel Klötgen	FRAU
Contributors	Salima Houta	FRAU
Contributors	Sofianna Menesidou	UBITECH
Contributors	Chrysostomos Symvoulidis	BYTE
Contributors	Stella Dimopoulou	BYTE
Contributors	Vincent Keunen	A7
Contributors	Lucie Keunen	A7
Contributors	Martin Marot	A7
Reviewers	Adrian Bradu	SIMAVI
Reviewers	Francesco Torelli	ENG

DRAFT

LOG TABLE

Version	Date	Change	Author	Partner
0.1	2021-05-25	V2 created as copy of v1, with first updates.	Gábor Bella	UNITN
0.2	2021-06-10	Updates and review of changes needed	Gábor Bella	UNITN
0.3	2021-06-23	Updated security specifications	Sofianna Menesidou	UBITECH
0.4	2021-06-28	Added RDD description	Niklas Haldorn	FRAU
0.5	2021-07-02	Added details on pseudo-identity generation	Stella Dimopoulou, Chrysostomos Symvoulidis	BYTE
0.6	2021-07-08	Reviewed and updated server-side interfaces	Simone Bocca, Alessio Zamboni	UNITN
0.7	2021-07-12	Security updates	Sofianna Menesidou	UBITECH

0.71	2021-07-12	Added comments	Stefano Dalmiani	FTGM
0.8	2021-07-13	Added and reviewed details on pseudonymisation and anonymisation	Gábor Bella, Stella Dimopoulou	UNITN, BYTE
0.9	2021-07-15	Added technical details on API descriptions	Simone Bocca, Alessio Zamboni	UNITN
0.91	2021-07-17	Finalised security specs	Sofianna Menesidou	UBITECH
0.92	2021-07-20	Prepared document for internal review	Gábor Bella	UNITN
0.93	2021-07-28	Internal review	Adrian Bradu	SIMAVI
0.94	2021-07-30	Quality check	Argyro Mavrogiorgou	UPRC
0.95	2021-08-02	Gabor Bella, Sofianna Menesidou, Athanasios Giannetsos	Last updates after QC and reviews	UNITN, UBITECH
Vfinal	2021-08-03	Final tech revision and version for submission	Francesco Torelli Laura Pucci	ENG

DRAFT

ACRONYMS

Acronym	Term	Definition
AES	Advanced Encryption Standard	AES is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology.
CA	Certificate Authority	An institution that issues digital certificates.
CN	Central Node	A node of the Research Network (a server) that stores published research studies and provides a central access point to S-EHR Apps for retrieving the descriptions of research studies.
-	Citizen	Any person potentially participating in a research study and having the minimal technical means to do so, i.e. the S-EHR App installed on their smartphone.
-	Client	The (public or private) legal entity who has ordered the research study and is paying for it.
CRC	Coordinating Research Centre	A medical research centre that initiates a particular research study and is in charge of defining it and carrying it out.
eIDAS	electronic IDentification, Authentication and trust Services	eIDAS is an EU regulation on electronic identification and trust services for electronic transactions in the European Single Market. Technically speaking, eIDAS infrastructure has been set up to connect the EU countries' national eID schemes to allow a person to authenticate in their home EU country when getting access to services provided by an eIDAS-enabled Service Provider in another EU country.
PI of the Research Centre	Principal Investigator of a Research Centre	The researcher (person) in charge of the citizens enrolled for a specific study at a RC.
PI of the Study	Principal Investigator of the Study	The researcher (person) in charge of a specific study at the CRC.
PP	Pseudonym Provider	An institution that generates and provides pseudonyms as a service.
RDD	Research Definition Document	A document written in a formal, computer-processable language that describes the research datasets to be retrieved from citizens' EHRs, enrolment and exit criteria, as well as related metadata.
RDDI	RDD Interface	Application Programming Interface allowing the exchange of RDDs between the Central Node and the S-EHR App.
RDS	Research Data Sharing	Acronym of the Research Data Sharing Protocol, the protocol covered by this deliverable
RDSI	RDS Interface	Application Programming Interface allowing the exchange of consent

DRAFT

		and health data between the S-EHR App and Research Centres.
RN	Research Network	The network of research centres and technical nodes that implement the Protocol.
RRC	Reference Research Centre (of a citizen)	A research centre participating in a given study that is a reference point for a specific citizen. The citizen sends health data to it for the duration of the study and the reference research centre is responsible for monitoring the citizen during the study.
SAML	Security Assertion Markup Language	SAML is an open standard for exchanging authentication and authorisation data between parties, such as between an identity provider and a service provider. SAML is an XML-based markup language for security assertions (statements that service providers use to make access-control decisions).

DRAFT

TABLE OF CONTENTS

DRAFT

DRAFT

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Scope of the Document

The overarching goal of the Research Data Sharing Protocol (in the following: *the Protocol*) is to specify a set of remote APIs and constraints on their usage that provide the technical means to citizens for the sharing of their health data for the purposes of cross-border medical research, in a cross-border interoperable manner. The particularity of this protocol, as opposed to current practices in medical research, is that it puts the citizens in full control of the sharing of their data: after explicit consent, data are retrieved directly from a citizen's mobile device in an anonymized manner.

From a technical point of view, the Protocol is almost peer to peer and decentralised, in particular the citizen shares the health data only with specific research centres. Only the metadata that describes the ongoing research studies are centralised. Moreover, the Protocol does not tie the citizens and research centres to specific software products, but specifies just the APIs and constraints that the interacting software systems must support and satisfy.

The Protocol addresses the giving and revocation of consent, the enrolment into specific research studies, the verification of enrolment criteria, as well as the retrieval and transfer of relevant health data. The Protocol defines the human and automated actors, the operations, and the communication channels and interfaces involved in these processes.

1.2 Intended Audience

This deliverable is intended primarily for a technical audience, interested in implementing the Protocol described in the document, or in understanding how data collection for cross-border medical research can be carried out with a direct involvement of citizens and their mobile-based health records. A certain familiarity of the medical research preparation process and of the challenges of cross-border data collection is useful for the understanding of this deliverable. Furthermore, the reader is expected to be familiar with certain other standards, formats, and specifications designed or used within the InteropEHRate project, such as the FHIR standard [FHIR], the InteropEHRate Architecture [D2.6] and the profiles for EHR interoperability [D2.8].

1.3 Structure of the Document

Section 2 provides a high-level overview of the Protocol. Section 3 defines the high-level architecture, interfaces of the typical Research Network that implements the Protocol, and the principal datatypes they use. Section 4 provides a brief summary of the contents of the machine-interpretable Research Definition Document. Section 5 describes the rationale and approaches used to anonymize and pseudonymize the health data shared by citizens. Section 6 describes the security aspects of the Protocol. Section 7 provides process definitions that clarify the interactions of the system components and human actors, in the form of activity and sequence diagrams. Section 8 describes the related work. Section 9, finally, provides conclusions.

DRAFT

1.4 Differences with respect to previous versions of the deliverable

With respect to the v1 of this deliverable, the Protocol maintains its principal architectural assumptions and communication process. It presents incremental additions as follows:

- in the new section 4, a summary of the Research Definition Document contents for information purposes (the RDD contents are specified in detail in [D2.9]);
- support for research questionnaires has been added throughout the document;
- the specifications of the Certification Authority, Pseudonym Provider, and eIDAS interfaces and the connections to them were removed, as they are outside of the scope of the Protocol (they are not standard interfaces and/or are defined in deliverables [D3.4] and [D3.10]);
- Citizen data anonymisation has been described in more detail in section 5.

In addition, the architecture of the Protocol has been extended to foresee the possibility of a direct communication between a research centre and the hospital of the citizen participating in a research study, in case not all data can be directly obtained from the Citizen's mobile device. The rationale for this addition is described more in detail in section 3. As the requirement for this feature was formulated by the time of finishing the current deliverable, not all technical details regarding the direct download of citizen health data from a hospital have been completely defined in the current version. For this reason, we foresee an addendum to D4.9 which will contain the missing technical details.

DRAFT

DRAFT

2 PROTOCOL OVERVIEW

This section provides an informal, high-level description of the goals, scope, participants, and processes covered by the Protocol. For a general high-level storytelling of how the Protocol is used to collect data for research studies[1], the reader is referred to the relevant section of deliverable [D2.2]. For simplicity, the specification sometimes refers to health data produced by hospitals, but it actually applies to citizen's health data produced by any other healthcare organisation.

2.1 Goals and Scope

The Research Data Sharing Protocol addresses the general problem of collecting health data for medical research directly from citizens, possibly involving citizens from multiple European countries. The motivations underlying the solution presented here are (1) to give more control to citizens over the use of their health data for research purposes; (2) to allow citizens to participate in research studies also remotely through their mobile devices; enable cross-border data collection in a way that involves citizens more directly in the decisions regarding the sharing of their data. This is achieved through a novel approach that retrieves data directly from the electronic health records stored on citizens' mobile devices. Citizens have complete control over their data as they can give or decline consent for data sharing on a per-study basis, and be informed of precisely what data are used by a given study.

In order to respond to the numerous technical challenges underlying such an approach, the Protocol brings novel solutions as well as relying on existing results from inside and outside the InteropEHRate project. It deals with the heterogeneity of cross-border data through relying on interoperable data representations, such as the *Interoperability Profile* defined by the InteropEHRate project [D2.8]. It relies on automatic data queries and on the checking of eligibility criteria inside the mobile device. It addresses privacy requirements by in-device or in-hospital data anonymisation, depending on data or usage constraints. It ensures the security of data transmission between mobile devices and research centres by relying on state-of-the-art encryption techniques. It provides a formal framework for consensual data sharing through digital signatures.

For simplicity, in the rest of this document we will assume that the mobile device of the citizen is a smartphone, but the Protocol actually applies to any mobile device of a citizen able to run a suitable mobile application supporting smart health data, such as the *S-EHR App* proposed by the InteropEHRate project.

[1] Note that in D2.2 a research study or its description are also called "research protocol". To avoid ambiguities, in the present document, the term "protocol" is used only to refer to the communication protocol for research health data sharing.

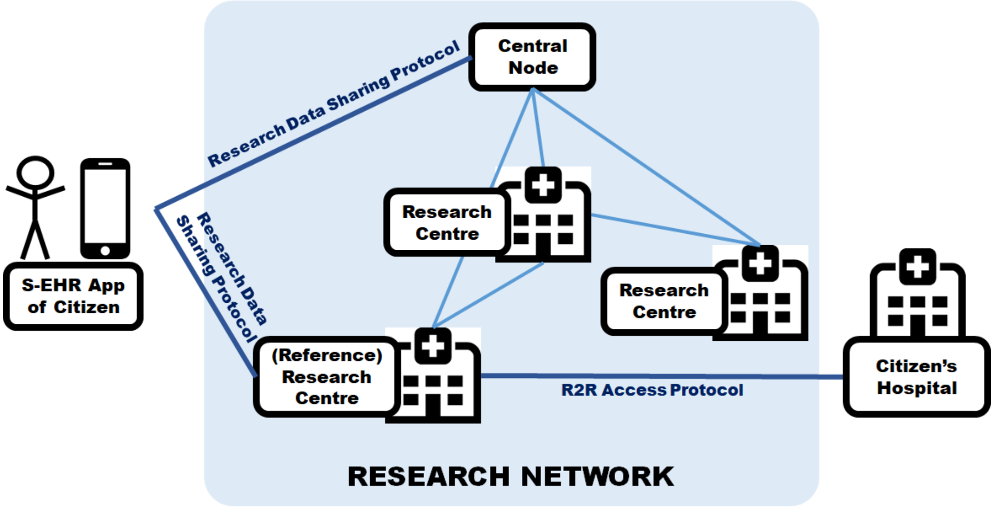


Figure 1 - High-level overview of the entities involved in the Research Data Sharing Protocol

Figure 1 shows a simple schematic diagram of the main components of a research data sharing scenario as assumed by the Protocol. The setup consists of (a) patients in possession of electronic health records stored on their mobile devices (the S-EHR App in the picture); (b) a *Research Network* that consists of interconnected *Research Centres* as well as a *Central Node*; and finally (c) a Hospital possibly outside of the Research Network that holds some of the Citizen's health data. In the case of a so-called *multi-centric research study*, multiple research centres may simultaneously collect data for the same study, each citizen being formally "attached" to a single research centre that becomes his or her *Reference Research Centre* (RRC) for that specific research study. The role of the Central Node is to store the formal, machine-interpretable definitions of research studies in the form of *Research Definition Documents* (RDD), and to provide these documents for download by mobile devices holding electronic health records. The S-EHR App then interprets the RDD and, in case the citizen is eligible and is willing to participate in the research study, retrieves relevant data from the citizen's health records and transmits them to the RRC in a fully secure and privacy-aware manner. The *Research Data Sharing Protocol* specifies the communication modalities between the S-EHR App and, on the one hand, the Central Node and, on the other hand, the citizen's Reference Research Centre. In case a piece of health data necessary for research is not directly downloadable from the Citizen's mobile device---because it is too big, then it is only present there as a reference as opposed to the actual dataset (such as a large media file), or it cannot be properly anonymized inside the mobile device for research purposes---the Protocol allows the Reference Research Centre to download the data directly from the Citizen's hospital, provided that the Citizen gives his/her consent to do so. This operation is realized through the *R2R Access* ("remote-to-research access") protocol, very similar to the *R2D Access* protocol defined by the InteropEHRate project in [D4.3].

2.2 Actors and Systems

The only human actors explicitly involved in the Protocol are *citizens.* A **citizen** is any person potentially participating in a research study with his/her health data, and having the minimal technical means to do so, i.e. a S-EHR App installed on his/her smartphone (or any other personal mobile device).

While not directly involved in any of the interactions in the scope of the Protocol, the following actors are participants of the overall research definition and data collection process and the semantics of some of the

DRAFT

exchanged data is related to them. Moreover they are responsible for some of the systems involved in the protocol.

- **Principal Investigator (PI) of the Study:** the researcher (person) in charge of a specific study, including its formal definition. The PI of the Study produces the Research Definition Document and has it published on the Central Node of the Research Network.
- **Principal Investigator (PI) of a Research Centre (RC):** the researcher (person) in charge of the patients enrolled for a specific study at a RC. The PI of the RC monitors the process of patient enrolment and the retrieval of their data.
- **Central Node Administrator:** a single person in charge of overseeing at the Central Node the publishing of new research studies on the Research Network.

These actors intervene through the following systems:

- **S-EHR App.** An application installed on the citizen's smartphone (or another type of supported personal mobile device) that stores and manages the citizen's health records, and is in charge of executing the Protocol implementation. It must fulfil the constraints specified by deliverable [D3.2].
- **Central Node (CN).** A node of the Research Network (a server) that stores published Research Definition Documents as defined by their respective PIs, and provides a central access point to S-EHR Apps for retrieving them.
- **Research Centre Information System.** The information system of a RC participating in a given study. It collects data shared by a set of citizens who are officially attached to this centre for the duration of the study.
- **Citizen's Hospital.** A hospital holding (some of) the citizen's health data. Upon request from the Research Centre Information System and permission from the citizen, the hospital can provide additional anonymised health data for a research study, in case such data is not obtainable through the S-EHR App.

The aforementioned actors also interact with the following security-related systems:

- **Certificate Authority (CA).** A trusted organisation that offers credential management services by issuing, certifying and removing digital certificates and the corresponding public keys linked to the long-term identity of their owners.
- **Pseudonym Provider (PP).** A trusted organisation that is responsible for the pseudonym management of the short-term anonymous credentials [1609.2-2016].

2.3 Data Exchanged

The following are the main kinds of data the exchange of which is covered by the Protocol:

- **Research Definition Documents:** structured documents formally describing research studies, including enrolment and exit criteria, data queries, a human-readable description of the study, a questionnaire to be shown to the citizen, and other study-related metadata.
- **Pseudonymized health data for research:** citizen health data, produced by hospitals or directly by the citizen, queried from the mobile device, pseudonymized/anonymized, and sent to a research centre.

DRAFT

- **Anonymized/pseudonymized health data not obtainable from the citizen’s mobile device:** such data is retrieved by a Research Centre through direct access to the citizen’s hospital.
- **Questionnaire responses:** the answers provided by the citizen to a study-specific questionnaire sent to them as part of the data collection process.
- **Digitally signed consent:** a formal agreement between a citizen and a research centre about the participation of a citizen to a research study, or his/her withdrawal from it.
- **Enrolment and exit notifications:** messages indicating the successful enrolment of a citizen into a study, or his/her leaving of the study.

For the representation of health data, as well as queries and criteria, the Protocol adopts the FHIR standard [FHIR], as does the entire InteropEHRate project. This design choice allows the retrieval of health data from citizens’ S-EHRs directly, without requiring further data conversion mechanisms. Beyond FHIR itself, the Protocol requires the data contained in S-EHRs to conform to the InteropEHRate FHIR profiles specified by the deliverable [D2.9], in order to ensure that cross-border data collection leads to meaningful results.

2.4 Processes

The execution of a research study, from its initial proposal by a researcher until its closure and archival, is a long and complex process that can last years, even for retrospective studies where medical data are readily available. Typically, the entire process involves the following macro-steps:

1. Pre-acceptance (GO / NO-GO)
2. *Formulation of requests to execute a given research study (as a formal research description)*
3. Approvals from the Ethical Committee as well as w.r.t. feasibility
4. Setting up of research environment
5. *Setting up the cohort, including citizen consent*
6. *Retrieval of data*
7. Preparation and linkage of datasets
8. Data analysis for the research experiment
9. Control of access to results
10. Archival of experiment and results
11. Closure

DRAFT

Addressing all of the macro-steps above is out of the scope of the InteropEHRate project and of the Protocol itself. The Protocol’s focus, instead, is the way in which medical data are retrieved directly from citizens’ smartphones, with all the necessary handling of consent, privacy, and security aspects of the operation. For this reason, the Protocol only covers the macro-steps relevant to these operations (in italics above), namely:

- **Formulation of request:** only to the extent that the research study is defined in the form of a formal, machine-processable RDD document. The Protocol does not cover *how* the RDD is created, but it does rely on the RDD in the operations it defines. The precise format of the RDD is defined in a separate deliverable on the InteropEHRate Interoperability Profiles [D2.9].
- **Setting up the cohort:** this covers the verification of enrolment criteria, as well as gathering citizen consent. Citizens are provided with the possibility of subscribing and being enrolled into specific research studies, as well as withdrawing from them.

- **Retrieval of data:** the citizens' data and questionnaire responses, if applicable, are transferred from the mobile devices (or in specific cases from the citizens' hospitals) to their respective RRCs.

Accordingly, the Protocol consists of the following macro-steps or phases:

1. **OPT-IN:** the Citizen opts in to be invited in research studies in general.
2. **ENROLLMENT:** the consenting Citizen is enrolled into a specific study.
3. **DATA RETRIEVAL:** relevant health data and questionnaire responses are retrieved from the Citizen's phone or from his/her hospital.
4. **WITHDRAWAL:** the Citizen decides to withdraw from providing further data to a given study.
5. **OPT-OUT:** the Citizen decides to opt out from a given study or from all current and future studies.

DRAFT

3 ARCHITECTURE AND INTERFACES

The figure below shows the main software systems, their exposed remote APIs, and their corresponding human users (actors) whose actions and communication are covered by the Protocol.

The protocol defines and exploits, as shown in Figure 2, the following APIs:

- **Research Data Sharing Interface (RDS):** REST API offered by the Research Centre Information System, allowing any S-EHR App to communicate the consent for a specific research study, receive enrolment-related information, and share citizen health data .
- **Research Definition Document Interface (RDD):** REST API offered by the Central Node, allowing the S-EHR App to download Research Definition Documents.
- **Remote-to-Research Access (R2RAccess):** REST API offered by the citizen's hospital for the retrieval of (typically very large) anonymous health data for research purposes.

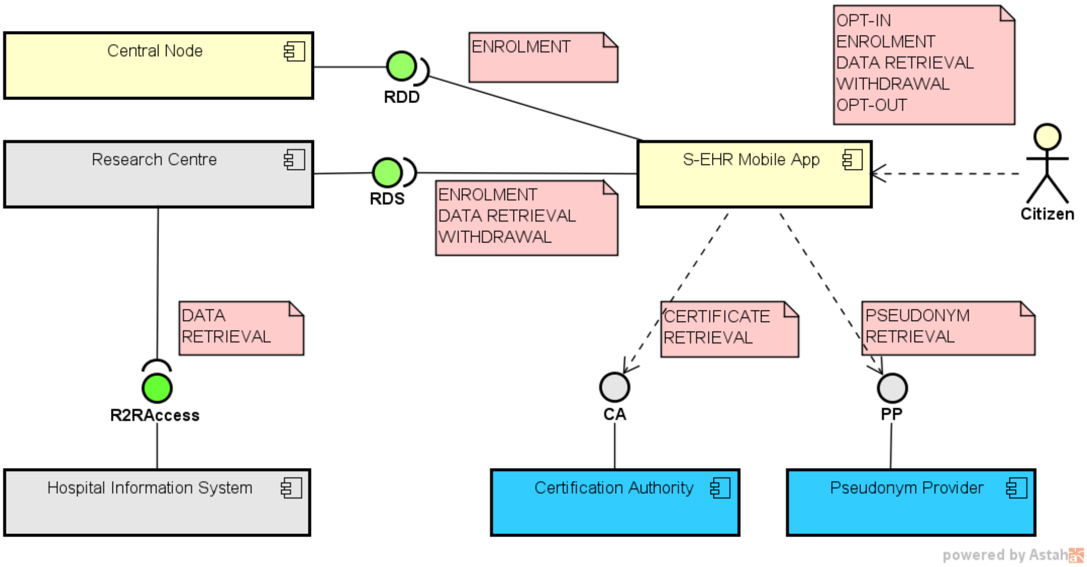


Figure 2 - Systems, actors, and communication channels of the Protocol

The meaning of colours in the figure is the following:

- blue for standard legacy interfaces and systems (used but not defined by the Protocol);
- grey for information systems of existing healthcare organisations and existing APIs;
- yellow for new systems defined in this deliverable;
- green for new interfaces defined in this deliverable.

3.1 Human-Computer Interfaces and Use Cases

This section describes the user interfaces that are required by the Protocol, from a high-level functional perspective of use cases. The Protocol covers the interactions of the Citizen with the Research Network. The Protocol does not specify how this human interaction happens, in particular, it does not require the usage of specific local APIs for executing them but specifies only how the input and output of these human interactions are related to input and output of the remote APIs specified by the Protocol. Other user interactions with the mentioned systems are possible, but they are not covered by the specification of the Protocol because they do not constraint the usage of the Protocol APIs.

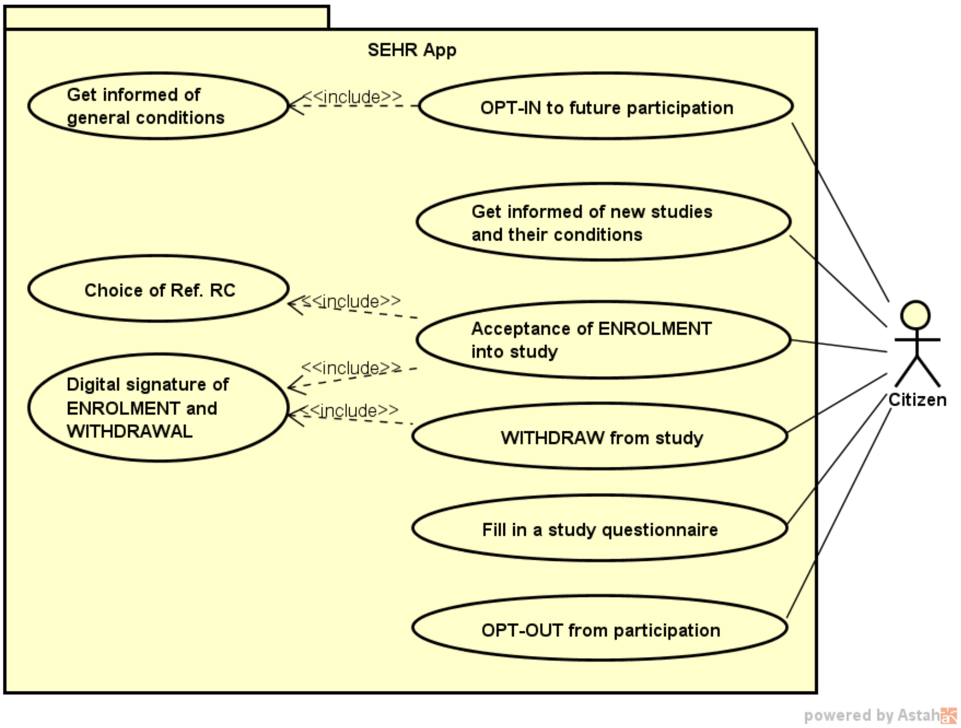

SEHR App
Get informed of general conditions
<<include>>
OPT-IN to future participation
Get informed of new studies and their conditions
Choice of Ref. RC
<<include>>
Acceptance of ENROLMENT into study
<<include>>
Digital signature of ENROLMENT and WITHDRAWAL
<<include>>
WITHDRAW from study
Fill in a study questionnaire
OPT-OUT from participation
Citizen
powered by Astah

Figure 3 - Use case diagram for the interaction of the Citizen with the S-EHR App

DRAFT

- **OPT-IN to future participation:** the Citizen sets his/her status on the smartphone as “interested” in participating in future studies. Before doing so, the Citizen is informed of what this entails (namely, the silent verification of enrolment criteria on his/her phone by accessing his/her health data, without sharing any citizen data with third parties). This allows the phone to retrieve regularly information about studies.
- **Get informed of new studies and their conditions:** the Citizen is informed about every study for which his/her health data meet the eligibility criteria, including the purpose and details of the study, the data collected, etc.
- **Acceptance of ENROLLMENT into study:** the Citizen formally accepts to participate in a given study.
- **Choice of Reference Research Centre:** as part of accepting the enrolment into a study, the Citizen chooses a reference research centre (RRC) from a list of possibilities corresponding to his/her geographical region of stay.
- **Digital signature of enrolment and of withdrawal:** for both enrolling into a study and withdrawing from it, a formal contract needs to be signed between the Citizen and his/her Reference Research Centre. These contracts are digitally signed by the Citizen, requiring his/her explicit participation.
- **Filling in a study questionnaire:** as part of providing health data for a given research study, the Citizen may be asked to fill in a questionnaire through the S-EHR App at the beginning of the study.
- **WITHDRAWAL from study:** the Citizen formally signals the decision to stop sending data for a given study.
- **OPT-OUT from participation:** the Citizen sets his/her status on the smartphone as “not interested” anymore in participating in future studies.

3.2 Remote APIs Defined by the Protocol

This section provides specifications for the endpoints of the three major interfaces defined by the Protocol:

- RDD, for communication between the Central Node and the S-EHR App;
- RDS, for communication between the Research Centre and the S-EHR App;
- R2RAccess and R2DAccessDICOM, for communication between the Research Centre and the Citizen's healthcare organisation.

Interfaces used but not specified by the Protocol (e.g. Certification Authority, eIDAS, and Pseudonym Provider API calls for security mechanisms) are not presented here as their specifications are provided as standards by third parties or within other InteropEHRate deliverables [D3.4], [D3.6] and [D3.10].

3.2.1 RDD - Central Node

The Central Node provides the services exposed through the *Research Dataset Definition Interface* (RDD). RDD is a RESTful interface.

Operation getOpenStudies

DRAFT

Property	Value
API Name	**getOpenStudies**
API Description	Allows any caller (but primarily a S-EHR App through a RESTful API call) to retrieve a digitally signed list of studies having the citizen enrolment period not yet expired, in other words, the studies open to any citizen who wants to participate. The RDD list which is retrieved through this service, contains all the RDDs published on the Research Network (using the publishStudy operation), which allow the enrolment of citizens (the enrolment period declared in the RDD is open).
Preconditions	In order to be able to check the digital signature in the return value, the caller needs to have access to the public key of the Central Node.
Caller	S-EHR App.
Return Value	FHIR Bundle
HTTP Method	GET
Header Params	● Content-Type: application/fhir+json
URL	http://<BASE_ADDR>/getOpenStudies?startDate=<DATE>
Input Parameters	● startDate: (optional) date value specified in ISO-8601 format (short form).

Return Value	A JSON object defined as follows:
	{
	"RDD": <FHIR-RDD-Bundle>,
	"signature": <digital-signature>,
	"certificate": <certificate>
	}
	where:

	● <FHIR-RDD-Bundle> is the FHIR RDD Bundle described in [D2.9]; ● <digital-signature> is the digital signature *string* used to sign the communication using the current endpoint; ● <certificate> is the Central Node certificate *string*, used to certify the source of the information retrieved through the current endpoint.
HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorisation is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR endpoint. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal error, the request could not be processed.
Example call	GET http://<BASE_ADDR>/getOpenStudies?startDate=20200615

3.2.2 RDS - Research Centre Information System

The Research Centre Information System provides the services exposed through the *Research Data Sharing Interface* (RDS). RDS is a RESTful interface. RDS offers the operation listed in the following table.

RDS Endpoint	Description
sendEnrollmentConsent	Send the Citizen’s electronically signed consent of enrolling into a specific study. The consent also includes the newly generated study-specific pseudonym or pseudo-identity, as well as the S-EHR App ID. The receiving RC checks the signature validity of the signedConsent, signs and returns the contract signed by both parties.
sendExitNotification	Send a notification that the Citizen is exiting a study. If the RRC fails to satisfy the call, a corresponding RESTful API Error is returned.
sendHealthData	Allows a S-EHR App to send citizen health data to the RRC. The receiving RC verifies and decrypts the encrypted and signed payload *healthData* and retrieves the FHIR bundle contained within. If the RRC fails to satisfy the call, a corresponding RESTful API Error is returned.
retrievePseudoIdentity	Allows a S-EHR App to receive a pseudo identity which has been generated at the RRC.

Table 1 - Methods of the RDSI Interface

DRAFT

Operation sendEnrollmentConsent

Property	Value
API Name	**sendEnrollmentConsent**
API Description	Send the Citizen’s electronically signed consent of enrolling into a specific study. The receiving RC checks the signature validity of the signed consent, signs and returns the contract signed by both parties.
Preconditions	● The S-EHR App of the Citizen must have verified the eligibility of the Citizen to participate in the study through checking the enrolment criteria. ● The S-EHR App must have access to the Citizen’s private key in order to sign the consent. ● The S-EHR App must have generated a pseudonym or pseudo-identity to be used in the study, which conforms to the RDD of the study.
Caller	S-EHR App.
HTTP Method	POST
Header Params	● Content-Type: `application/fhir+json`
URL	`http://<BASE_ADDR>/sendEnrollmentConsent?studyID=<studyID>`

Input Parameters	URL params:
	● studyID: the ID of the study in which the Citizen is enrolling;
	The POST body content is a JSON file defined as follows:
	`{`
	`    "consent": <signed-consent>,`
	`    "citizen-pseudo": <citizen-pseudo>,`
	`    "certificate": <citizen-certificate>,`
	`    "enrollment-criteria-data": <enrollment-criteria-data>,`
	`    "sehrapp-id": <sehr-app-id>`
	`}`
	where:
	● `<signed-consent>`: a digitally signed document (encoded in Base64) containing the Citizen’s consent to participate in the study;
	● `<citizen-pseudo>`: pseudonym or pseudo-identity generated for the Citizen;
	● `<citizen-certificate>`: certificate of the Citizen issued by a Certification Authority and sent to the RC so that it can verify the digital signature;
	● `<enrollment-criteria-data>`: encrypted (JSON) data values corresponding to the enrolment criteria, so that the RC can cross-check their validity (see below this table for an example of enrolment-criteria-data JSON object);
	● `<sehr-app-id>`: identifier of the S-EHR App product and instance, for traceability of the S-EHR App product being used).

DRAFT

Return Value	The response body is a JSON file defined as follows:
	`{`
	`"signed-contract": <signed-contract>,`
	`"certificate": <rc-certificate>`
	`}`
	where:
	● `<signed-contract>` is the consent contract where the Research Centre has added its own digital signature, and which is now signed by both parties;
	● `<rc-certificate>` is the certificate of the Research Centre that certifies the authenticity of the digital signature.

HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorisation is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR endpoint. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal error, the request could not be processed.

Exceptions	The call's exceptions returned are added as text messages within the HTTP response body which is defined as follows:
	`{`
	`"timestamp":<timestamp>,`
	`"status":<http-status-code>,`
	`"error":<code-description>,`
	`"message":<exception message>,`
	`"path":<request-path>`
	`}`
	where:
	● `<timestamp>` is the response timestamp;
	● `<http-status-code>` is one of the http codes listed in the previous row;
	● `<code-description>` is the description of http code;
	● `<request-path>` is the http request's URL;
	● `<exception message>` is one of the following text messages:
	○ invalid content (study ID, pseudo-identity, consent form);
	○ digital signature cannot be verified;
	○ enrolment criteria not met.

Example call	`POST http://<BASE_ADDR>/sendEnrollmentConsent?studyID=123`

Example input parameter	An example of `enrollment-criteria-data` provided as input parameter:
	`{`
	`"age": 23,`
	`"gender": "female"`
	`}`

DRAFT

Operation sendExitNotification

Property	Value
API Name	**sendExitNotification**
API Description	Send a notification that the Citizen is exiting a study due to the exit criteria being met **or upon explicit withdrawal by the Citizen.**
Preconditions	● The Citizen must have enrolled into the study previously.
Caller	S-EHR App.
HTTP Method	POST
Header Params	● Content-Type: `application/fhir+json`
URL	`http://<BASE_ADDR>/sendExitNotification?studyID=<studyID>`

Input Parameters	URL params
	● studyID: the ID of the study in which the Citizen is enrolling;
	The POST body content is a JSON file defined as follow:
	`{`
	`"citizen-pseudo": <citizen-pseudo>,`
	`"reason": <reason>,`
	`"reason-text": <reason-text>,`
	`"citizen-signature": <citizen-signature>`
	`}`
	where:
	● `<citizen-pseudo>`: the study-specific pseudonym or pseudo-identity of the Citizen;
	● `<reason>`: either the string “WITHDRAWAL” or “DROPOUT”, the first one meaning a voluntary withdrawal by the Citizen while the second is the consequence of the exit criteria being met;
	● `<reason-text>`: in the case of DROPOUT, a text string explaining the reason for the exit (e.g. “blood pressure < 120”);
	● `<citizen-signature>`: digital signature confirming the exit.

Return Value	None
HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorisation is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR endpoint. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal

DRAFT

	error, the request could not be processed.

Exceptions	The call's exceptions returned are added as text messages within the HTTP response body which is defined as follows:
	`{`
	`    "timestamp":<timestamp>,`
	`    "status":<http-status-code>,`
	`    "error":<code-description>,`
	`    "message":<exception message>,`
	`    "path":<request-path>`
	`}`
	where:
	● `<timestamp>` : response timestamp;
	● `<http-status-code>` : one of the http codes listed in the previous row;
	● `<code-description>` : description of http code;
	● `<request-path>` : http request's URL;
	● `<exception message>` : one of the following text messages:
	○ invalid content (study ID, pseudo-identity, reason, signature);
	○ citizen not enrolled and thus cannot exit.

Example call	POST http://<BASE_ADDR>/sendExitNotification?studyID=123

DRAFT

Operation sendHealthData

Property	**Value**
API Name	**sendHealthData**
API Description	Allows a S-EHR App to send citizen health data to a Research Centre. The receiving RC verifies and decrypts the encrypted and signed payload healthData and retrieves the FHIR bundle contained within.
Preconditions	● The Citizen must have enrolled into the study previously. ● The S-EHR App must have access to the Citizen's private key to encrypt the health data, and the called Research Centre must have access to the Citizen's public key to be able to decrypt it.
Caller	S-EHR App
FHIR Resource involved	Bundle
HTTP Method	POST
Header Params	● Content-Type: application/fhir+json
URL	http://<BASE_ADDR>/sendHealthData?studyID=<studyID>
Input Parameters	URL params ● studyID: the ID of the study in which the Citizen is enrolling; The POST body content is a JSON file defined as follows:

	`{`
	`"citizen-pseudo": <citizen-pseudo>,`
	`"health-data": <health-data>`
	`}`
	where:
	● `<citizen-pseudo>`: the study-specific pseudonym or pseudo-identity of the Citizen;
	● `<health-data>`: a FHIR bundle containing the health data (resources, attributes, values) necessary for the study, in an encrypted form, as well as the responses to research questionnaire(s) provided by the Citizen if available

Return Value	None
HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorisation is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR endpoint. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal error, the request could not be processed.

Exceptions	The call's exceptions returned are added as text messages within the HTTP response body which is defined as follows:
	`{`
	`"timestamp":<timestamp>,`
	`"status":<http-status-code>,`
	`"error":<code-description>,`
	`"message":<exception message>,`
	`"path":<request-path>`
	`}`
	where:
	● `<timestamp>` : response timestamp;
	● `<http-status-code>` : one of the http codes listed in the previous row;
	● `<code-description>` : description of http code;
	● `<request-path>` : http request's URL;
	● `<exception message>` : the following text message:
	○ invalid content (study ID, pseudo-identity, healthData)

Example call	`POST http://<BASE_ADDR>/sendHealthData?studyID=123`

DRAFT

Operation retrievePseudoIdentity (for pseudo-identity-based studies)

Property	Value
API Name	**retrievePseudoIdentity**
API Description	Allows a S-EHR App to receive a pseudo-identity which has been generated at the RRC.
Preconditions	● The S-EHR App must have checked that the Citizen's data fulfils the enrolment criteria for the study, and that the Citizen consents to participating in the study. ● The RDD specifies that the RRC must generate a pseudo-identity
Caller	S-EHR App
HTTP Method	GET
Header Params	● Content-Type: `application/json`
URL	`http://<BASE_ADDR>/retrievePseudoIdentity?studyID=<studyID>`
Client Params	URL params ● studyID: the ID of the study in which the Citizen is enrolling.

Return Value	A JSON object defined as follow:
	`{`
	`  "message": <message>,`
	`  "pseudo-identity": <pseudo-id>,`
	`  "status": <code>`
	`}`
	where:
	● `<message>`: a message that inform about the pseudo-identity generation (success or not);
	● `<pseudo-id>`: the pseudo-identity represented as a string;
	● `<code>`: the HTTP status of the request.

HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorisation is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR endpoint. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal error, the request could not be processed.

DRAFT

Exceptions	The call’s exceptions returned are added as text messages within the HTTP response body which is defined as follows:
	`{`
	`    "timestamp":<timestamp>,`
	`    "status":<http-status-code>,`
	`    "error":<code-description>,`
	`    "message":<exception message>,`
	`    "path":<request-path>`
	`}`
	where:
	● `<timestamp>` : response timestamp;
	● `<http-status-code>` : one of the http codes listed in the previous row;
	● `<code-description>` : description of http code;
	● `<request-path>` : http request’s URL;
	● `<exception message>` : one of the following text messages:
	○ invalid content (study ID);
	○ if the study does not require a pseudo-identity (i.e. it requires a pseudonym obtained through a different channel).

Example call	`GET http://<BASE_ADDR>/retrievePseudoIdentity?studyID=123`

3.2.3 R2R-Access - Remote-to-Research Access

This communication protocol is an instantiation of the *R2D Access* data transfer protocol defined by the InteropEHRate project in deliverable [D4.3]. Originally, the purpose of R2D Access is to allow the transfer of health data from a healthcare organisation towards a mobile device. In the scope of research data sharing, the role of the target device is played by the citizen’s Reference Research Centre.

A detailed description of the R2R-Access endpoints used for research data sharing will be provided as an addendum to this deliverable.

3.3 FHIR Binding of API Datatypes

The technical binding of the API datatypes is based on the data model as described in [D2.9], where several FHIR Implementation Guides are described. The Implementation Guide for Research Data Sharing defines the data model for the management of research related studies.

Several constraints apply to the technical binding that go beyond the definition of a data model:

- An overarching and unique study identifier related to a specific research study must be provided and used with the data exchange related to that study.
- Healthcare data related to a study must not contain any identifying data such as a citizen’s name, instead the provided healthcare data should be anonymized or pseudonymized.
- Identifiers used to uniquely identify a citizen, either on a device, or in a data sharing environment, must be replaced with pseudo-identifiers (or pseudonyms) related to a citizen in the context of a specific study. This way, the identity of the citizen is protected.

DRAFT

- The provision of research data is the result of the application of the enrolment criteria, specifying a citizen cohort with certain features, and the data selection criteria, defining exactly which information about a citizen are used to compile the research data. Thus, only the data defined by the data selection criteria of citizens matching the cohort specification are provided. The selected research data must be anonymized and all identifiers must be replaced with pseudo-identifiers (or pseudonyms) as described above in order to protect the citizen's identity.
- The enrolment criteria and the data selection criteria are provided as part of a research study.
- A trusted organisation, such as the research center, may be in possession of the citizen's unique identifiers and the citizen's pseudo-identifiers and is able to re-identify a citizen by managing their correspondence.

The following table shows the API parameter binding.

DRAFT

No	API parameter	technical binding (see D2.8)		description
		implementation guide	HL7 FHIR resources / profiles (Cardinality)	
1	FHIR-RDD-Bundle	Implementation Guide for Research Data Sharing	- Research Study (1..N) - Cohort (1..N) - Data Set Definition (1..1) - Reference Research Centres (1..N)	A digital signed list of currently open studies. It consists of a FHIR bundle containing a combination of the resources shown in this table and profiled according to the Implementation Guide for Research Data Sharing [D2.8].
2	studyID	Implementation Guide for Research Data Sharing	Research Study . identifier (1..1)	A ResearchStudy object which contains the attribute identifier.
3	citizenIdentification	Implementation guide for Cross Border Data Exchange	Patient	Identifying attributes of a citizen should not be transmitted together with the corresponding pseudonyms used for a specific study. Therefore, the Patient profile is used.
4	citizenPseudo	Implementation Guide for Research Data Sharing	- Research Subject . pseudoID (1..1) - Research Subject . patient . identifier	The API parameter citizenPseudo contains either a pseudonym or a pseudo-identity of the citizen, and is transmitted based on the ResearchSubject profile. If the connection between a pseudo and a patient must be

				transmitted, the different identifiers can be transmitted within the same resource.
5	signedConsent	Implementation Guide for Research Data Sharing	Research Subject . Citizen Consent (1..1)	The API parameter signedConsent is based on the Consent profile referenced by the Research Subject profile.
6	sehrAppId	Implementation Guide for Research Data Sharing	- Device . identifier (1..1) - Device . patient . id (1..1)	The device identifier attribute is used in order to specify the potentially identifying S-EHR App Id. A non-anonymous patient identifier is also provided in order to allow the assignment to an identified citizen.
7	Reason	Implementation Guide for Research Data Sharing	profile/extension on Research Subject . status	
8	HealthData	Implementation guide for Cross Border Data Exchange	All	A digitally signed and encrypted bundle of FHIR resources according to the specified implementation guide. All identifying information has been removed or replaced in the context of the research study. Instead, the research subject's pseudo id and anonymized demographic information is contained. The bundle also contains a FHIR resource that represents a study-related questionnaire and the responses given to it by the citizen.

Table 2 - Technical bindings of the API parameters

DRAFT

4 RESEARCH DEFINITION DOCUMENT CONTENT SUMMARY

The detailed specifications of the RDD content and structure are provided in [D2.9]. Here we offer only a summary of the RDD contents in order to provide a high-level understanding to the reader of how research studies are formally described, destined both to inform citizens of the purpose of the study and to be interpreted automatically by the S-EHR App on their mobile devices. This summary provides a short description for each resource that is part of the Research Definition Document and the relation between those resources.

RDSI Endpoint	Description
ResearchStudy	The *ResearchStudy* resource is the entry point for the Research Definition Document. It contains general information about the study like the purpose, the title and a description. The ResearchStudy also contains references to a Cohort, that contains information about the entry and exit criteria of the study, a list of ResearchCenters, that contain information about the locations taking part in the study, a DataSetDefinition, that defines the data requested from the participants of the study, and Questionnaires. Further information is provided by extensions that define the level of anonymisation required for data sent as a port of the study and the enrolment period.
Cohort	Exactly one *Cohort* resource is referenced in each ResearchStudy and describes the entry and exit criteria for the study. The cohort can contain any number of characteristics that are either marked as include or exclude. A characteristic additionally contains the kind of characteristic, the value it should have and during which time it should have had this value. To be eligible to participate in the study the patient must have all characteristics marked with include and none of the characteristics marked with exclude.
ResearchLocation	Multiple *ResearchLocations* can be referenced in each ResearchStudy and describe a location that fulfils a role in the study. As the ResearchLocation only describes a location and not an organisation there can be multiple ResearchLocation resources for a single organisation, for example if the organisation has multiple laboratories taking part in the study. The ResearchLocation must contain the name, address and contact information for this location. It also defines the role this location plays in the study and provides a reference to an technical EndPoint, that can be used for electronic requests to the location.
DataSetDefinition	Exactly one *DataSetDefinition* extension is part of each ResearchStudy and contains the DataRequirements for the study. The DataSetDefinition contains a DataRequirement for each piece of data the study wants to collect. The DataRequirement contains the exact data that should be sent and a time period during which it should be sent. By starting this period in the past data from past observations and by setting the end date to a future date data from future observations can be requested. The DataRequirement can also contain a frequency extension that describes with what frequency updated versions of the data should

DRAFT

	be sent.
Questionnaires	Multiple *Questionnaire* extensions can be part of each ResearchStudy and contain a Reference to a Questionnaire and information about the timeframe during which the questionnaire has to be completed. The Questionnaire that is referenced in this extension is provided by the author of the study and is used to collect additional information about the participant. The extension also provides the deadline for the Questionnaire and the frequency with which the app should remind the participant to fill out the Questionnaire.

Table 3 - Summary contents of the RDD

DRAFT

5 ANONYMIZATION AND PSEUDONYMIZATION

DRAFT

This section describes the pseudonymisation and anonymisation of health data shared by citizens for research purposes. These mechanisms are necessary because citizens' S-EHR Apps contain personally identifiable information and sensitive data (i.e. health data). In order to ensure that participation in research studies remains anonymous, both pseudonymisation and anonymisation need to be performed so that the data sent to researchers does not include any unique identifier that could lead to revealing the identity of the Citizen.

The difference between anonymisation and pseudonymisation is, while in neither case should the identity of a participating citizen be revealed, pseudonymisation still allows the identification of the citizen. This possibility is foreseen only within the context of the Reference Research Centre and in exceptional circumstances, such as upon the discovery of a severe illness of a citizen as part of the research process, requiring the researcher to be able to contact the citizen immediately. The RRC can map the pseudo-identity to a citizen's unique identifiers, or can request the mapping of the pseudonym to be done by a third party, such as a Pseudonym Provider.

In the context of InteropEHRate, depending on the mechanism by which the pseudo-identifier of a citizen is generated, one of two variants of pseudonymisation will be used: (1) *pseudo-identity-based* or (2) *pseudonym-based*, both of which are presented in detail in section 4.1 below. The Principal Investigator of the Study is responsible for choosing the variant to use, as well as whether a simple anonymisation of certain datasets is sufficient. These details are also set inside the Research Definition Document (RDD), which can include policies for all cases. In case of pseudonymisation, all personal information found in the requested data will be replaced with a pseudo-identity or pseudonym, whereas in case of anonymisation all personal information will be removed so that it can be impossible for someone to lead to the identification of a citizen.

The pseudo-anonymisation of certain types of content---such as imagery where the patient identity is embedded into the bitmap, or unstructured documents such as PDF files---may be out of the reach of the capabilities of certain mobile devices. While we expect the storage and computing power of smartphones to continue to increase in the near future, the Protocol also foresees the need to perform pseudo-anonymisation outside of the mobile device. More precisely, as evoked in sections 2 and 3, the S-EHR App, and indirectly the Citizen, may provide consent to the Reference Research Centre to request anonymized content directly from the Citizen's hospital, in case such content is missing from the data transmitted by the S-EHR App but is made available for download by the hospital.

5.1 The Pseudo-Anonymisation Process

Operations related to pseudo-anonymisation are involved in multiple phases of the research data sharing process:

1. when the study is being defined: definition of privacy policies for the study;
2. during citizen enrolment: generation of pseudo-identities or pseudonyms for the Citizen;
3. during data retrieval: pseudo-anonymisation of research data, either by the Citizen's S-EHR App or by the Citizen's hospital.

Definition of pseudonymisation and anonymisation policies for the study. The Principal Investigator (PI) responsible for creating the Research Definition Document (RDD) for a research study, defines the pseudonymisation and anonymisation policies with respect to the datasets to be retrieved as part of the research study. This includes whether to use pseudonyms or pseudo-identities, how the pseudo-identifier should be formatted, as well as which datasets need to be pseudonymized and which need to be anonymized. [D2.8] contains details about the way requirements are defined inside the RDD.

Generation of pseudo-identities or pseudonyms. For each study and for each participating citizen, a new pseudo-identity or pseudonym needs to be generated in order uniquely to identify him/her without revealing his/her identity. The details of this process are given in Section 5.2 below.

Pseudo-anonymisation of data. Before health data is sent to a research centre, it needs to be pseudonymized or anonymized. Depending on the kind of data (structured, unstructured, text, image, video, etc.), the process may be very different and more or less complex technically. In consequence, anonymized data may need to be retrieved from the Citizen's hospital instead of the mobile device. Details are provided in Section 5.3 below.

5.2 Pseudo Generation

In the case of pseudonymisation, there are two ways to replace all personal information of the citizens, depending on the pseudonymisation policy adopted for the research study. Once the citizen gives consent to participating in the study, either a (more conventional) *pseudo-identity* or a (stronger but more rarely used) *pseudonym* [Camenisch 2017] will be created for the citizen. The pseudo-id will be generated by a pseudo-identity generation service at each Research Centre participating in the study, whereas the pseudonym will be generated by the Pseudonym Provider (PP). Depending on study-specific policies defined within the RDD and initially set by the PI of the study, the mechanism of either using pseudo-identities or pseudonyms is chosen.

5.2.1 Variant #1: Study-Specific Pseudo-Identities

The pseudo-identity is essentially a string generated after a pattern that consists of numbers and/or letters. It replaces all attributes, and in particular all direct and indirect identifiers, which are not included in the requested data and can lead to the identification of a citizen, for example the name and age. The pseudo-id consists of three parts:

[PREFIX] [INCREMENTED_NUMBER] [SUFFIX]

The prefix is an alphanumeric sequence that depends on the study, and is stated in the given RDD. The incremented number is increased every time a pseudo-id is created for a citizen for a specific study. The suffix is a random sequence, such as an SHA-256 alphanumeric string.

This method has its pros and cons: the advantage of using this method is that the pseudo-ids are human-readable, which corresponds to current well-established practices of hospitals and researchers. The disadvantage is that the pseudo-ids have limited randomness and they are more vulnerable to unauthorized de-identification, as opposed to the pseudonyms.

Following long-standing practices in medical research, pseudo-identities are generated by the Reference Research Centre for each enrolled citizen. The pseudo-identity is requested by the S-EHR App of the Citizen through a dedicated API call to the RRC in the enrolment phase. The RRC, in turn, uses a separate *pseudo-identity generation service* that may be implemented as an internal service of the RRC or, as an extra privacy measure, by a *trusted third party*. Traditionally, it is the PI of each study who maintains the mapping table between citizen identities and pseudo-identities, and who is responsible for re-identification in case

DRAFT

of emergency (such as a serious illness discovered during the research process). The Protocol does not specify the precise mechanism or architecture through which pseudo-identifiers should be managed. Reliance on a trusted third party in order to manage (store, re-identify, delete) pseudo-identities is a practice that enhances the preservation of privacy. The pseudo-identity is returned from the RRC to the S-EHR App that stores it at least until the end of the data retrieval period of the study, for purposes of in-phone pseudonymisation.

In case of an emergency scenario, the Principal Investigator is the one who is responsible for the re-identification of the citizen. More specifically, the PI of the study maintains a mapping table which contains the pseudo-identity and the personal information of each citizen. In this way, the PI can be led to the identity of the citizen in a non-automatic way and only if this is necessary for a specific purpose.

5.2.2 Variant #2: Short-Term Pseudonyms

Pseudonyms are certificates that are only valid if they are signed by a root CA and only for a short time [Eckhoff2011]. As a precondition to requesting the issuance of pseudonyms, at the time of retrieving his/her electronic health record from a hospital, the citizen must have acquired his/her certificate [eHDSI2021] and retrieve an anonymous assertion that can later be used to anonymously authenticate him/herself to the PP as a legitimate entity. The issuance of this anonymous assertion can be supported by any identity provider used to check the identity of the citizen when requesting the issuance of a certificate, from the CA, or the CA itself as part of the common criteria defined for the certificate policy management [LABIOD2018]. In our case since eIDAS is also used as the identity provider we also consider the interaction between the CA and the eIDAS for verifying the ID of the citizens [D3.6]. Thus the CA upon request of a certificate issuance of a user interacts with eIDAS to check the ID of the citizens; if verification is successful the eIDAS SAML response contains also this anonymous token (currently supported by default by the standard eIDAS response data model). This token is used for the Citizen's authentication to the Pseudonym Provider, in order to receive a high-entropy pseudonym. Entropy provides the measure of the uncertainty to identify the citizen that is participating in a study among a set of citizens. Overall, the issuance of the certificate and the anonymous token is a precondition. Especially for the former this is also aligned to what has been described as a prerequisite in the enrolment phase (see section 7.3). Indeed, the eIDAS is one possible way for the CA to verify the ID of the citizen and get the anonymous assertion as part of the eIDAS SAML response. However, this anonymous assertion can also be created by the CA or another trusted entity (other than eIDAS) that is used to check the ID of the citizen [eHealth2021].

After the pseudonym has been received, the Citizen can send his/her signed consent (signed with his/her private key) to the RRC and, finally, can start sharing (anonymously signed) data. The latter also contains a blind signature so that the RRC can always verify that received data originates from a user that has already provided a signed consent. The PI of the RRC will be able to get access to the mapping of pseudonyms with the IDs of the citizens that acquired them in order to be also able to perform quick pseudonym resolution. No other actor from the RRC---or from anywhere else---will be able to link data back to the users/citizens, unless in a situation of emergency. The actors that can request pseudonym resolution (beyond the PI of an RRC that directly has access to the mapping of pseudonyms to the IDs of citizens) will have to prove both their identity and the reason why this request needs to take place (e.g. medical emergency).

5.3 Anonymisation Operations

After the execution of the research data query in the S-EHR App, but before forwarding the data to the Research Centre, the data retrieved needs to be pseudo-anonymized within the S-EHR app before being

DRAFT

sent to the researcher. Beyond the removal of identifying information such as name or address, irrelevant information not explicitly requested by the researcher also has to be excluded.

The main challenges of the anonymisation operation therefore are:

- how to know which datasets (files, data attributes) need to be searched for identifying information;
- how to know the different types of identifying information that may appear inside data;
- how to find identifying information within datasets with a high enough precision and recall, making sure that (ideally) all occurrences of such information are found and that only identifying information is eliminated at the end.

The three challenges above are of considerably different complexity depending on the format of the data in question, in particular with respect to whether it is in a standard structured or in an unstructured form. In the following, we thus distinguish these two cases that require diverging approaches.

5.3.1 Anonymisation of Structured FHIR Data

Thanks to the fact that the Citizen's S-EHR App contains his/her health data in a standard FHIR-based structured format, the anonymisation operation over such data is a relatively straightforward operation that is feasible even with low computational resources, including the mobile device itself. An important requisite, however, is to maintain up-to-date information on the version of the FHIR standard in use (including its possible extensions, as defined in the Interoperability Profiles [D2.9]) and its potentially identifying attributes. In other words, the anonymisation logic needs to maintain:

- for each FHIR resource that may potentially appear in the S-EHR App, the list of attributes that either entirely *consist of* identifying information or *may contain* such information;
- the *kind of identifying information* contained in such attributes (person name, address, birth date, social security number, etc.).

Maintaining the list of attributes is far from trivial as the FHIR standard consists of thousands of attributes, it is extensible, and---as any other standard data schema---it keeps evolving. Likewise, maintaining an exhaustive list of the kinds of identifying information that may potentially appear inside attribute values is difficult due to the well-known inherent complexity of the de-identification and re-identification problem. For instance, individual attributes may not be identifying in themselves but can be so in combination, and it is very difficult---if not impossible---to foresee all possible (involuntary or malicious) re-identification mechanisms. Therefore, it is important to adopt standard, state-of-the-art approaches to anonymisation and to state very clearly and transparently to the Citizen the approach being applied and its potential limitations.

The anonymisation process for structured data will be described in detail in a separate deliverable [D6.8].

5.3.2 Anonymisation of Unstructured Data

As of today, a large proportion of electronic health record data provided by hospitals remains in unstructured form. This includes reports in free text or PDF format (e.g. discharge report) as well as medical imagery including graphs as well as still and moving imagery (e.g. X-ray, electrocardiogram, echocardiogram). Anonymisation of images is relevant as such images often contain text embedded on the pixel level by the device used to produce them that often contains identifying information, such as the name of the patient.

Many solutions exist for the de-identification of unstructured data. As a general rule, the more specific the solution, the more straightforward it is to implement it. For instance, it is relatively easy to cover up

DRAFT

embedded text in images produced by a specific device of a specific healthcare organisation using simple manually defined rules. Several DICOM image management tools are capable of handling such specific cases. Providing a general mechanism that finds identifying text in all kinds of images is, on the other hand, extremely difficult. Likewise, the discharge report of a given hospital may be easily analyzed in a focused manner, but a generic analysis method for any PDF document written in any language, using any local convention, is on a different level of complexity.

For this reason, the most sensible approach to the anonymisation of unstructured data is to execute it as close to its source as possible, i.e. within the healthcare organisation that has produced it. Firstly, the healthcare organisation knows best the data and therefore is the most likely to be able to define the anonymisation process with sufficient accuracy. Secondly, the execution of generic---and thus very complex, resource-hungry, time-consuming, and error-prone---anonymisation operations inside the S-EHR App on a mobile device remains unrealistic as of today.

The precise anonymisation operations to be applied are outside of the scope of this deliverable. Anonymisation of structured content will be covered in [D6.8]. However, the Research Data Sharing Protocol does address the issue of anonymisation feasibility on an architectural and communicational level: it foresees a dedicated communication channel between the research centre collecting research data from citizens and the hospital capable of anonymizing the data it has produced. While the exact process for retrieving anonymized data from hospitals will be provided in an addendum to this deliverable (as explained in Section 1), below we provide the main steps of the process.

1. When data is retrieved from the mobile device to be sent to the Reference Research Centre, unstructured data objects that need to be anonymized before being sent to the RRC are identified by the S-EHR App. Instead of sending these objects to the RRC, the S-EHR App provides references (resource identifiers) that allow their retrieval from the originating hospital. For each object, the S-EHR App also provides a *Request Authorisation Token* that testifies that the Citizen has given consent to the RRC to retrieve the objects from the hospital.
2. Upon reception of the references and tokens, the RRC makes a request to the hospital(s) holding the data.
3. The hospital performs the anonymisation on the fly (taking as much time as needed) and exposes the result for subsequent download by the RRC.

DRAFT

6 PROTOCOL SECURITY

This section explains how the requirements (see Table 4) for the security of citizen health data and research study metadata exchange are met by the Research Data Sharing Protocol. In particular, it covers the security of messages between:

- the S-EHR App and the Central Node in the context of downloading RDDs;
- the S-EHR App and the Reference Research Centre in the context of data sharing;
- the Reference Research Centre and the hospital providing anonymized data in the context of data retrieval from the hospital through *R2R-Access*.

Table 7 lists high-level user requirements for the Protocol that have security implications. Based on these requirements, the main security requirements to be considered are:

- the encryption of the communication channel between the S-EHR App and the Reference Research Centre in order to achieve confidentiality of the shared medical data;
- the authenticity and integrity of the medical information;
- the authenticity and integrity of RDDs;
- mutual authentication between the Citizen and the Reference Research Center with the requirement of privacy-preserving authentication from the RRC side.

Public keys and certificates. Establishing a Public Key Infrastructure (PKI) is one of the most important tasks in security concerning communication over the internet. To be able to do so, a trusted third-party Certificate Authority (CA) is required to be in place. The role of the CA is a) to issue certificates, b) to confirm the identity of a certificate owner, and c) to provide proof that the certificate is valid. For multiple mechanisms and security concepts to work, the above requirements must be fulfilled. However, until now there is no standard API to be used to issue a certificate.

Issuing of certificates. In order to apply all the security needs in the Protocol, a necessary bootstrap phase exists in order for all the participants to issue their certificates from a Certificate Authority, and pseudonym certificates from a Pseudonym Provider. This phase is out of the scope of the protocol, however it is mandatory for the successful completion of the security steps. Section 3.2 provides all the interfaces of the Research Data Sharing Protocol, including the CA APIs as well as the necessary security parameters, such as *encryptedBundle*.

Encryption of shared data. To guarantee data confidentiality, an authenticated key agreement protocol will be used to securely exchange a symmetric session and AES256 for the actual encryption. The Station-to-Station (STS) [STS1992] protocol is a cryptographic key agreement scheme based on classic Diffie–Hellman, and provides mutual key and entity authentication. Unlike the classic Diffie–Hellman, which is not secure against a man-in-the-middle attack, this protocol assumes that the parties have signature keys, which are used to sign messages, thereby providing security against man-in-the-middle attacks. In the context of the research scenario, the STS scheme will be used securely to establish a symmetric encryption key. The authentication form the reference Research Center side is basically encapsulated in the consensus to enrolment as defined in Section 3, where the reference Research Center needs to authenticate the citizen in a privacy-preserving manner through the appropriate use of pseudonyms.

DRAFT

Transport-layer security. Last but not least, apart from the application-level security, *Transport Layer Security v1.2* must be enabled at the transport layer in both communication channels, the RDSI and RDDI to protect from data breaches and Distributed Denial of Service (DDoS) attacks. In a nutshell, TLS is a protocol which provides privacy between communicating applications and their users, or between communicating services. When a server and client communicate, well-configured TLS ensures that no third party can eavesdrop or tamper with any message.

#id	User requirement title	Security Requirements
#69	Non-repudiable data provenance tracking	Non repudiation
#70	Integrity of medical information	Integrity
#86	Digital signature by Reference Research Centre of Citizen's consent	Authenticity
#87	Citizen's digital signature of consent to share health data for a given study	Authenticity
#88	Citizen's digital revocation of consent to share health data for a given study	N/A
#130	Pseudoidentity restricted to single research protocol	Privacy
#150	Identification and authorisation of organisations and researchers accessing to IRS	Authentication & Authorisation

Table 4 - Security Requirements

DRAFT

6.1 Security Prerequisites

The correct execution of the Protocol supposes that the following prerequisites are respected regarding credentials:

- the Central Node, the S-EHR App, and all Reference Research Centres have retrieved their certificates from a central Certification Authority, as described in the context of other protocols in [D3.6];
- upon downloading health data, the Citizen has obtained an eIDAS authentication token, to be used to connect to the Pseudonym Provider (see [D3.4] for details);
- upon installation, the S-EHR App has downloaded the public key of the Central Node along with its connection address (URI).

6.2 Security of the RDS Interface

Below we present in detail the operations performed by Research Centres as well as the S-EHR App in order to establish the security of communication performed by means of the RDS interface. The table below maps the security operations to the Protocol steps in which they are executed, as described in Section 6.

Purpose: Confidentiality of the communication channel between the S-EHR App and the reference Research Center, Integrity and Authenticity of the shared medical data, Mutual Authentication between the Citizen and the reference Research Center.

Actors and components: Citizen, S-EHR App, reference Research Centre, Pseudonym Provider, Principle Investigator of the RRC, Certificate Authority.

Preconditions: All actors have installed the necessary credentials and certificates signed and verified by a CA. For the pseudonym-based variant of pseudonymisation, the Citizen must also be already authenticated to the trusted Pseudonym Provider using a valid eIDAS SAML assertion and retrieve the necessary assertion (e.g. anonymous certificate). TLSv1.2 or greater needs to be established.

Steps:

Step	Security Operation	Protocol Step (see Section 6)
1	The S-EHR App uses a Key Derivation Function (KDF) to derive a one-time key K that will be used for encrypted communication between the S-EHR App and the reference Research Centre. Then encrypts asymmetrically the generated key K with the public key of the reference Research Center.	ENROLLMENT Step 4
2	The S-EHR App sings the assertion that verifies that the Citizen is authenticated and acquired signed from the eIDAS Node, using the pseudo-identity/pseudonym that acts as a short-term signing key.	ENROLLMENT Step 5
3	The S-EHR App encrypts the double-signed assertion asymmetrically with the public key of the reference Research Center.	ENROLLMENT Step 5
4	The S-EHR App sends to the reference Research Center the concatenated encrypted signed assertion and the encrypted generated symmetric key.	ENROLLMENT Step 7
5	The reference Research Center receives the message, verifies the S-EHR App's signature, decrypts the encrypted key with its private key then verifies the signed assertion.	ENROLLMENT Step 7
6	The reference Research Center adds its own digital signature to the signed consent document using its private key, and encrypts it with the S-EHR App's public key, before sending the counter-signed consent document back to the Citizen.	ENROLLMENT Step 8
7	Upon reception of the counter-signed consent document, the Citizen decrypts it using his/her private key, and checks that it is signed by the reference Research Center using the Center's public key.	ENROLLMENT Step 8
8	The S-EHR App encrypts symmetrically and sings the anonymised data to be shared with the pseudo-identity or pseudonym.	DATA RETRIEVAL Step 5
9	The reference Research Centre receives the encrypted and anonymised health data and validates the signature. Then decrypts	DATA RETRIEVAL Step 6

DRAFT

	data using the established key.	
10	The Citizen withdraws his/her participation in a study, digitally signing a withdrawal message (contract) before sending it to the reference Research Center.	WITHDRAWAL Step 1

Table 5 - Mapping of RDS security operations to the Protocol steps in section 7

These security operations correspond to three major security requirements:

- **mutual authentication** between the S-EHR App and the Reference Research Centre before starting sharing the data;
- **confidentiality** of the shared data through encryption, so that unauthorized individuals cannot access or use them;
- **non-repudiation** of the data sharing contract or its withdrawal, through digital signatures.

6.3 Security of the RDD Interface

Below we present in detail the operations performed by the Central Node as well as the S-EHR App in order to establish the security of the communication performed by means of the RDD interface. The table below maps the security operations to the Protocol steps in which they are executed, as described in section 6.

Purpose: Integrity and Authenticity of RDDs, Non-repudiation of the Central Node signatory.

Actors and components: Citizen, S-EHR App, Central Node, Principal Investigator of the Study, Certificate Authority.

Preconditions: All actors should have installed the necessary credentials and certificates signed and verified by a CA. The Central Node should also acquire its own certificate signed by the trusted Certificate Authority. TLSv1.2 or greater should be used. The PI has already authenticated to the Central Node and published a digitally signed RDD.

Steps:

Step	Security Operation	Protocol Step (see Section 6)
1	The Central Node digitally signs the RDDs, to ensure that it is not accidentally altered or changed with her/his private key.	Before ENROLLMENT Step 1
2	The S-EHR App downloads the published RDDs from the Central Node along with their digital signatures and certificate.	ENROLLMENT Step 1
3	The S-EHR App checks the validity of each certificate and signature. If the validation is successful, the S-EHR App can be sure that the study was not altered (integrity), while the Central Node cannot deny having published the RDDs (authenticity, non-repudiation).	ENROLLMENT Step 2

Table 6 - Mapping of RDD security operations to the Protocol steps in section 7

DRAFT

The rationale behind these security aspects of RDD interface is to download the RDDs and to be sure that nothing has been altered during the download. Integrity and non-repudiation are necessary aspects of health data, documents, and reports. For this reason, InteropEHRate should comply with the Electronic Signatures Directive 1999/93/EC and EU Regulation 910/2014 of 23 July 2014 on electronic identification (eIDAS). Digital signature ensures authenticity and integrity of the RDDs, while at the same time prevents the Central Node signatory from being able to repudiate (deny) his involvement. Such properties make the adoption of digital signatures an integral part of the RDD interface.

6.4 Security of the R2R Access Interface

This section details the security specifications for the download of unstructured anonymized data from the Citizen's hospital to the RRC via the *R2R-Access* protocol. The contents of this section will be provided as part of a future addendum to this deliverable.

DRAFT

7 PROCESS DEFINITIONS

This section describes the communication process among the human actors and systems involved, as defined by the Protocol the sequence diagrams corresponding to each phase of the protocol. The Protocol is divided into phases as defined in section 2. For each phase, both a high-level activity diagram and a lower-level and more formal sequence diagram are provided. The diagrams focus on the interactions between components and thus contain little detail on operations internal to single components. Green colour marks the swimlanes and actions of human agents (Citizen), while light yellow represents actions by automated systems.

7.1 Prerequisite: Obtainment of Security Certificate

Before research data sharing can be initiated from the Citizen's mobile device, it is required that the Citizen should have downloaded an X.509 certificate (certified public-private key pair) from the Certificate Authority and installed it. While this operation is equally a prerequisite in other use cases of InteropEHRate and is therefore not strictly part of the Research Data Sharing Protocol, we present it here for purposes of completeness. More detail can be found in the deliverable [D3.6].

Actors and components: Citizen, S-EHR App, Certificate Authority.

Preconditions: The S-EHR App has been installed on the Citizen's mobile device.

Steps:

1. The S-EHR App checks if a certificate already exists on the device.
2. If no certificate is installed, the S-EHR App requests one from the CA by first generating a key pair and then by sending a certificate signing request (CSR)[2] to the CA. The CSR usually contains the signed public key for which the certificate should be issued, as well as identifying information. The most common format for CSRs is the *PKCS #10 specification*.
3. If the request is successfully verified, the CA issues a certificate that has been digitally signed using the private key of the CA Then, the CA returns the issued certificate as the response to the call.
4. The S-EHR App stores the certificate returned in the keystore of the device.

DRAFT

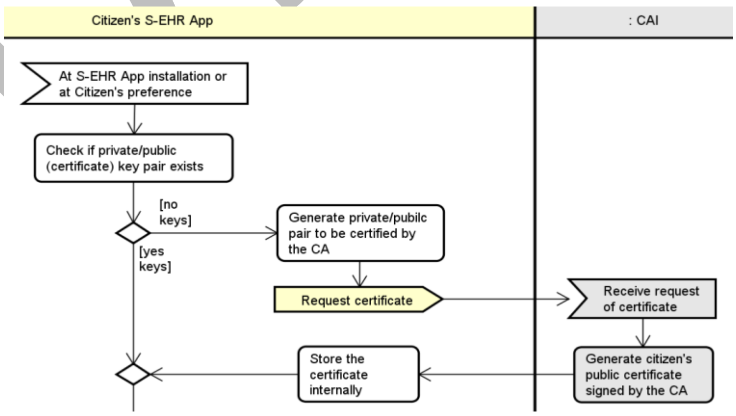

Citizen's S-EHR App
: CAI
At S-EHR App installation or at Citizen's preference
Check if private/public (certificate) key pair exists
[no keys]
[yes keys]
Generate private/pubilc pair to be certified by the CA
Request certificate
Receive request of certificate
Store the certificate internally
Generate citizen's public certificate signed by the CA

Figure 4 - High-level data flow of the security certificate retrieval

[2] https://en.wikipedia.org/wiki/Certificate_signing_request

7.2 OPT-IN phase

Purpose: In the OPT-IN phase, the Citizen signals the general intention of participating in research studies in the future. (S)he gives his/her consent to the S-EHR App on the phone regularly to poll the Research Network for new studies soliciting enrolment. The solicitation of the Citizen could possibly happen in the last stages of the installation process.

Actors and components: Citizen, S-EHR App.

Preconditions:

- The S-EHR App has been installed on the Citizen's mobile device.

Steps:

1. The S-EHR App solicits the Citizen for a general agreement to future participation in research studies.
2. The Citizen chooses either *yes* (opt-in) or *no* (opt-out). The S-EHR App may also allow the Citizen to postpone the decision.
3. The Citizen's answer is recorded in the S-EHR App. In case the Citizen decided to opt out, the S-EHR App does not send any other research-related solicitations to the Citizen in the future.

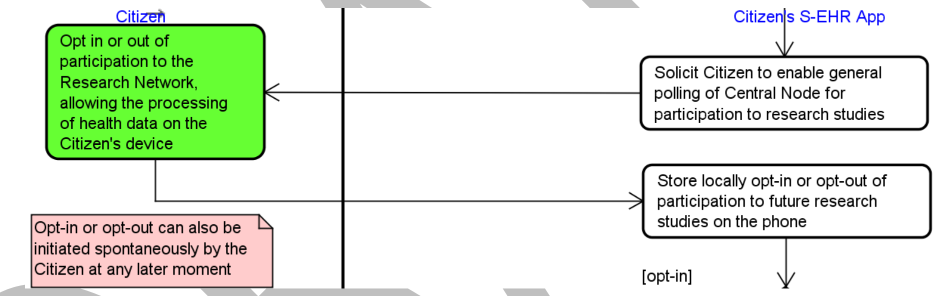

Citizen
Citizen's S-EHR App
Opt in or out of participation to the Research Network, allowing the processing of health data on the Citizen's device
Solicit Citizen to enable general polling of Central Node for participation to research studies
Store locally opt-in or opt-out of participation to future research studies on the phone
Opt-in or opt-out can also be initiated spontaneously by the Citizen at any later moment
[opt-in]

Figure 5 - High-level data flow of the OPT-IN phase

7.3 ENROLLMENT phase

Purpose: In the enrolment phase, for each newly published study, the S-EHR App evaluates whether the Citizen's health data matches the enrolment criteria, and if so, asks for the Citizen's consent to be enrolled in the study. Upon enrolment, a Reference Research Centre is assigned to the Citizen or selected by him/her from multiple research centres in case the study is multi-centric.

Actors and components: Citizen, S-EHR App, Reference RC, Central Node.

Preconditions:

- The Citizen has previously opted in for participating in studies.
- A study has been published on the Central Node.
- The Citizen has a S-EHR installed on his/her smartphone that contains health data in a formal, structured, and standardised representation, such as the one defined by the InteropEHRate Interoperability Profile [D2.8].
- The Citizen has previously obtained his/her security certificate from the Certificate Authority.

DRAFT

- For participation in pseudonym-based (and not pseudo-identity based) studies, an anonymous authentication token has been previously acquired by the Citizen, issued by the CA or any other Identity Provider leveraged for verifying the ID of the Citizen.

Steps:

1. If the Citizen has chosen to opt in to studies in the OPT-IN phase, the S-EHR App regularly (e.g. daily) polls the Central Node to retrieve the RDDs of currently open studies. The set of such RDDs is downloaded by the S-EHR App from the Central Node as a digitally signed document.
2. The S-EHR App checks the digital signature on the set of RDDs retrieved.
3. For each new study RDD retrieved, the S-EHR App silently extracts and verifies the enrolment Criteria with respect to the patient data of the Citizen. In case the patient data does not meet the enrolment Criteria, the RDD is silently deleted and no further action is taken with respect to it.
4. If the patient data meets the enrolment Criteria, the Citizen is solicited by the S-EHR App for his/her (digitally signed) consent to participate in the study, displaying to him/her the goals, purposes, and conditions of research and the data collected. In case the Citizen declines participation, the RDD is deleted and no further action is taken with respect to it.
5. Once the Citizen has agreed and digitally signed his/her consent, if the study is multi-centric, he/she is prompted to choose a Reference Research Centre (RRC) from the list of research centres contained in the RDD (typically a research centre close to his/her residence).
6. An anonymous study-specific identifier is generated to be used for data pseudonymisation in the data retrieval phase. Depending on the study definition within the RDD, either a pseudo-identity or a pseudonym is generated (see section 4):
    a. in case the study operates with pseudo-identities:
        i. the S-EHR App retrieves a pseudo-identity from the RRC chosen by the Citizen;
        ii. the consent to enrolment, digital signature, and S-EHR App identifier are then sent to the RRC chosen by the Citizen. The S-EHR App identifier is needed so that the RRC can contact the Citizen if necessary, such as if the Citizen's health data point to an important and so far undiagnosed pathology.
    b. in case the study operates with short-term pseudonyms:
        i. the S-EHR App retrieves a study-specific pseudonym, created by the Pseudonym Provider that represents the Citizen;
        ii. the consent to enrolment, digital signature, S-EHR App identifier, and pseudonym are then sent to the RRC. The S-EHR App identifier is needed so that the RRC can contact the Citizen if necessary, such as if the Citizen's health data point to an important and so far undiagnosed pathology.
7. The RRC receives this enrolment notification and stores it in its list of citizens enrolled into the study.
8. The RRC counter-signs the consent using digital signature, and sends back the counter-signed document to the S-EHR App, which saves it locally.

DRAFT

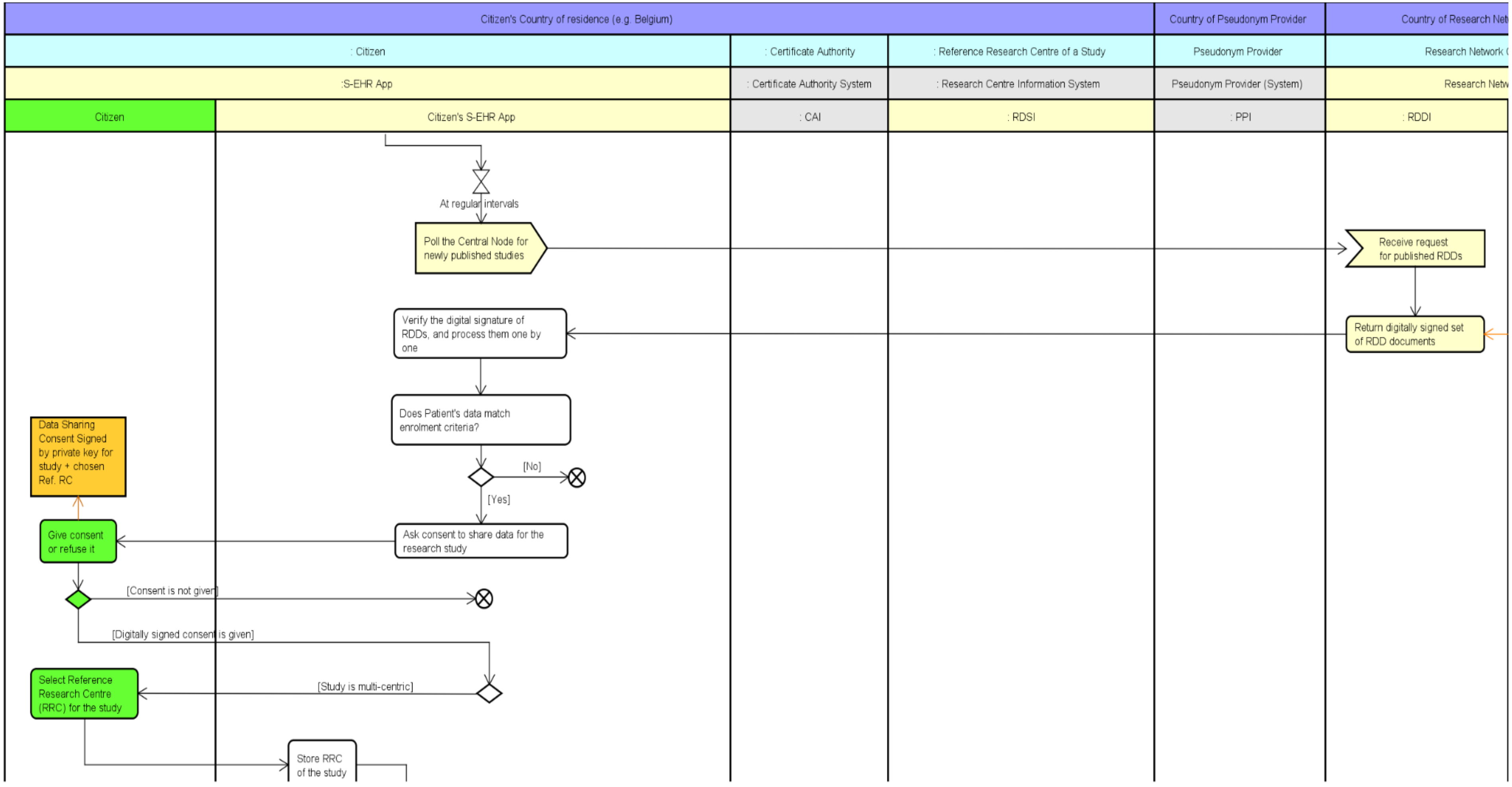

Citizen's Country of residence (e.g. Belgium)
Country of Pseudonym Provider
Country of Research Net
: Citizen
: Certificate Authority
: Reference Research Centre of a Study
Pseudonym Provider
Research Network (
:S-EHR App
: Certificate Authority System
: Research Centre Information System
Pseudonym Provider (System)
Research Netw
Citizen
Citizen's S-EHR App
: CAI
: RDSI
: PPI
: RDDI
At regular intervals
Poll the Central Node for newly published studies
Receive request for published RDDs
Return digitally signed set of RDD documents
Verify the digital signature of RDDs, and process them one by one
Does Patient's data match enrolment criteria?
[No]
[Yes]
Ask consent to share data for the research study
Data Sharing Consent Signed by private key for study + chosen Ref. RC
Give consent or refuse it
[Consent is not given]
[Digitally signed consent is given]
[Study is multi-centric]
Select Reference Research Centre (RRC) for the study
Store RRC of the study

Figure 6a - High-level data flow of the ENROLLMENT phase (beginning)

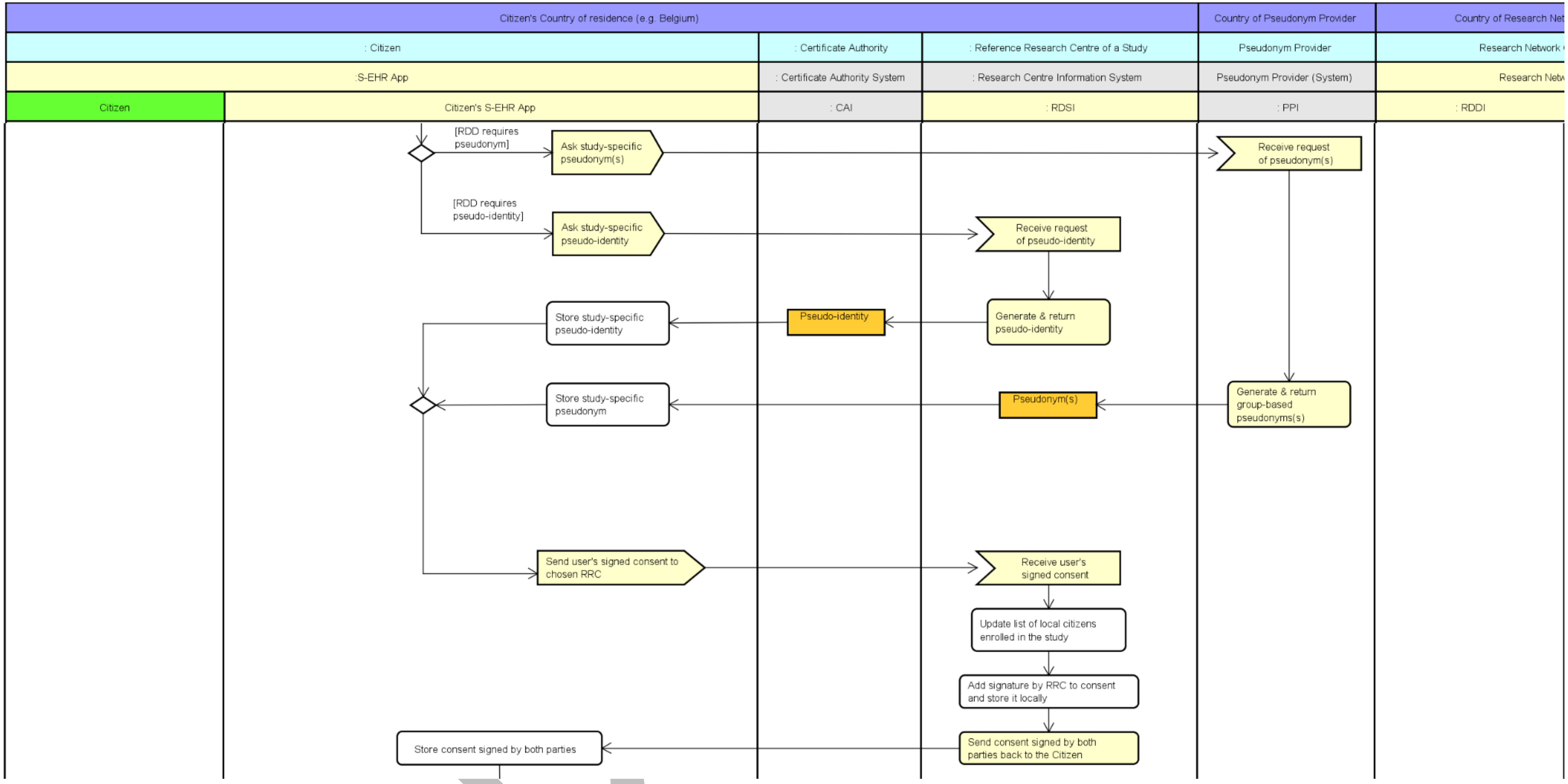
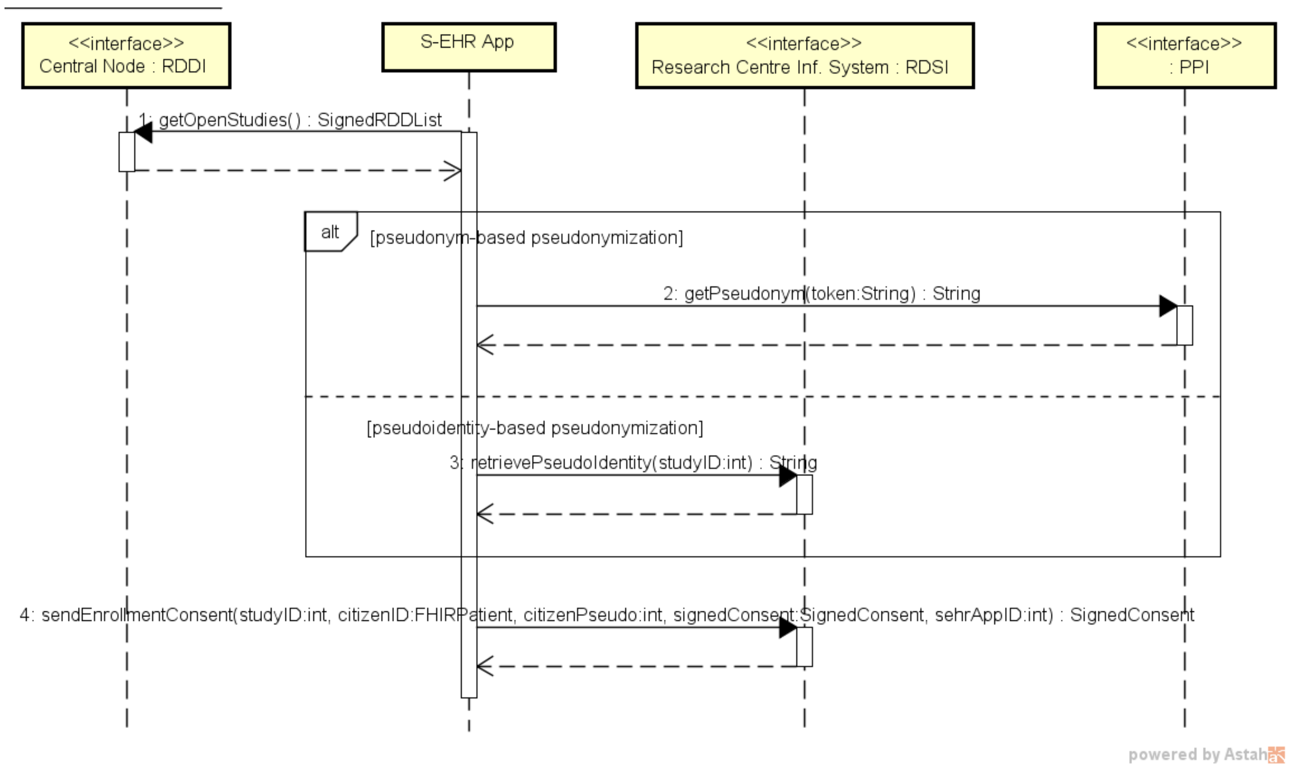

Citizen's Country of residence (e.g. Belgium)
Country of Pseudonym Provider
Country of Research Net
: Citizen
: Certificate Authority
: Reference Research Centre of a Study
Pseudonym Provider
Research Network
: S-EHR App
: Certificate Authority System
: Research Centre Information System
Pseudonym Provider (System)
Research Netw
Citizen
Citizen's S-EHR App
: CAI
: RDSI
: PPI
: RDDI
[RDD requires pseudonym]
Ask study-specific pseudonym(s)
Receive request of pseudonym(s)
[RDD requires pseudo-identity]
Ask study-specific pseudo-identity
Receive request of pseudo-identity
Store study-specific pseudo-identity
Pseudo-identity
Generate & return pseudo-identity
Store study-specific pseudonym
Pseudonym(s)
Generate & return group-based pseudonyms(s)
Send user's signed consent to chosen RRC
Receive user's signed consent
Update list of local citizens enrolled in the study
Add signature by RRC to consent and store it locally
Store consent signed by both parties
Send consent signed by both parties back to the Citizen

Figure 6b - High-level data flow of the ENROLLMENT phase (continued)



<<interface>>
Central Node : RDDI
S-EHR App
<<interface>>
Research Centre Inf. System : RDSI
<<interface>>
: PPI
1: getOpenStudies() : SignedRDDList
alt
[pseudonym-based pseudonymization]
2: getPseudonym(token:String) : String
[pseudoidentity-based pseudonymization]
3: retrievePseudoIdentity(studyID:int) : String
4: sendEnrollmentConsent(studyID:int, citizenID:FHIRPatient, citizenPseudo:int, signedConsent:SignedConsent, sehrAppID:int) : SignedConsent
powered by Astah

Figure 7 - Sequence diagram for the ENROLLMENT phase

7.4 DATA RETRIEVAL phase

Purpose: In the DATA RETRIEVAL phase, data relevant to the study is gathered from the Citizen's smartphone, either a single time (for retrospective studies) or repetitively (for prospective studies). After in-phone anonymisation, the data are sent to the Citizen's RRC.

Actors and components: S-EHR App, Reference RC.

Preconditions: The Citizen has been enrolled in the study.

Steps:

1. If there are one or more questionnaires associated with the study, the Citizen is presented with each questionnaire during the data retrieval period, at a time predefined by the study. For each questionnaire, while the Citizen is not forced to fill it in immediately, he/she will be regularly reminded by the S-EHR App of the need to contribute his/her answers. Filling in the questionnaires is mandatory: if the Citizen does not do so until the end of the Data Retrieval period, the consequence is that he/she will be excluded from the study by his/her Reference Research Centre. This exclusion is implicit and is not covered by the Protocol (there is no message sent by the RRC to the Citizen about the exclusion).
2. If the study is retrospective AND all necessary data are available, the following steps 2-6 will be executed only once. Otherwise, they will be executed periodically, as defined in the RDD. Within each data retrieval period, the S-EHR App regularly (e.g. daily) checks for new, updated data to be available, and executes steps 2-7 as soon as this is the case.
3. The S-EHR App verifies again if the Citizen meets the exit criteria. If the result is positive, an Exit Notification is sent to the RRC containing the reason (enrolment criteria negative or exit criteria

DRAFT

positive). Upon the reception of this message, the RRC updates the list of citizens enrolled into the study.

4. The S-EHR App retrieves the data to be sent to the RRC, building and executing the query based on the dataset definition included in the RDD.
5. The S-EHR App applies anonymisation and pseudonymisation to the data retrieved, based on the requirements included in the RDD.
6. The S-EHR App sends the anonymized/pseudonymized data to the RRC. Depending on policy/settings, the S-EHR App may notify the Citizen that data has been sent.
7. The RRC receives the data and verifies the identity of the sender.
8. The RRC verifies if the Citizen has not been in the meantime withdrawn from the study by the PI of the Study due to an exit event undetected by the S-EHR App (such as a hospitalisation). If this is the case, the RRC responds to the data retrieval with a “not enrolled anymore” message.
9. If the data received contains references to content necessary for the study stored at a remote hospital (as well as a corresponding Request Authorisation Token provided by the Citizen), the RRC retrieves the content from the hospital. This step, not yet depicted in Figure 7, will be detailed in an addendum to this deliverable.
10. If the data reception was completed successfully, a “success” message is returned as response to the S-EHR App.
11. At the end of the data retrieval period, upon the decision of the PI of the RRC, the RRC forwards the data collected from all citizens under its control to the CRC where the PI of the Study resides.

DRAFT

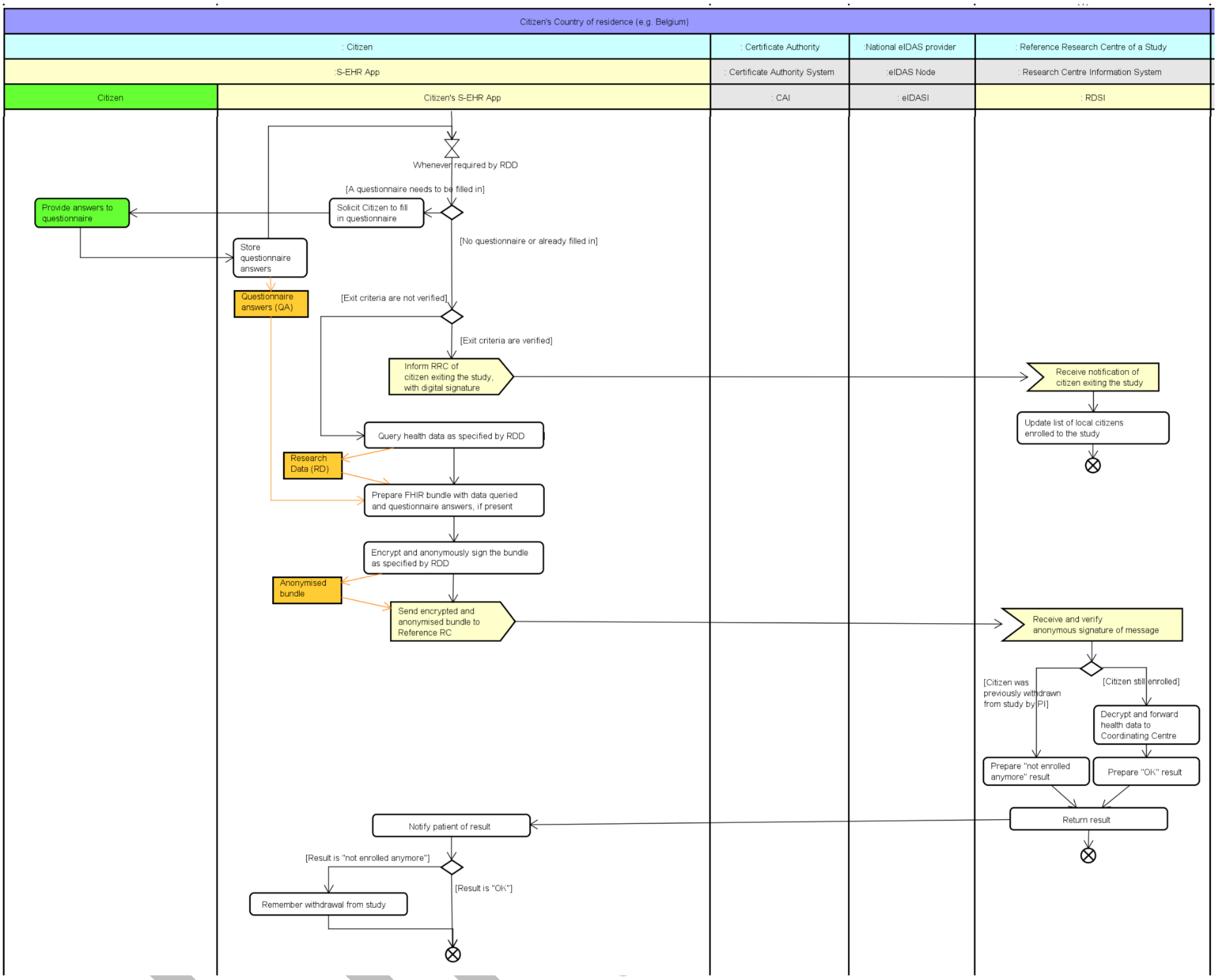

Citizen's Country of residence (e.g. Belgium)
: Citizen
: Certificate Authority
:National eIDAS provider
: Reference Research Centre of a Study
S-EHR App
: Certificate Authority System
:eIDAS Node
: Research Centre Information System
Citizen
Citizen's S-EHR App
: CAI
: eIDASI
: RDSI
Whenever required by RDD
[A questionnaire needs to be filled in]
Provide answers to questionnaire
Solicit Citizen to fill in questionnaire
[No questionnaire or already filled in]
Store questionnaire answers
Questionnaire answers (QA)
[Exit criteria are not verified]
[Exit criteria are verified]
Inform RRC of citizen exiting the study, with digital signature
Receive notification of citizen exiting the study
Update list of local citizens enrolled to the study
Query health data as specified by RDD
Research Data (RD)
Prepare FHIR bundle with data queried and questionnaire answers, if present
Encrypt and anonymously sign the bundle as specified by RDD
Anonymised bundle
Send encrypted and anonymised bundle to Reference RC
Receive and verify anonymous signature of message
[Citizen was previously withdrawn from study by PI]
[Citizen still enrolled]
Decrypt and forward health data to Coordinating Centre
Prepare "not enrolled anymore" result
Prepare "OK" result
Return result
Notify patient of result
[Result is "not enrolled anymore"]
[Result is "OK"]
Remember withdrawal from study

Figure 8 - High-level data flow of the DATA RETRIEVAL phase

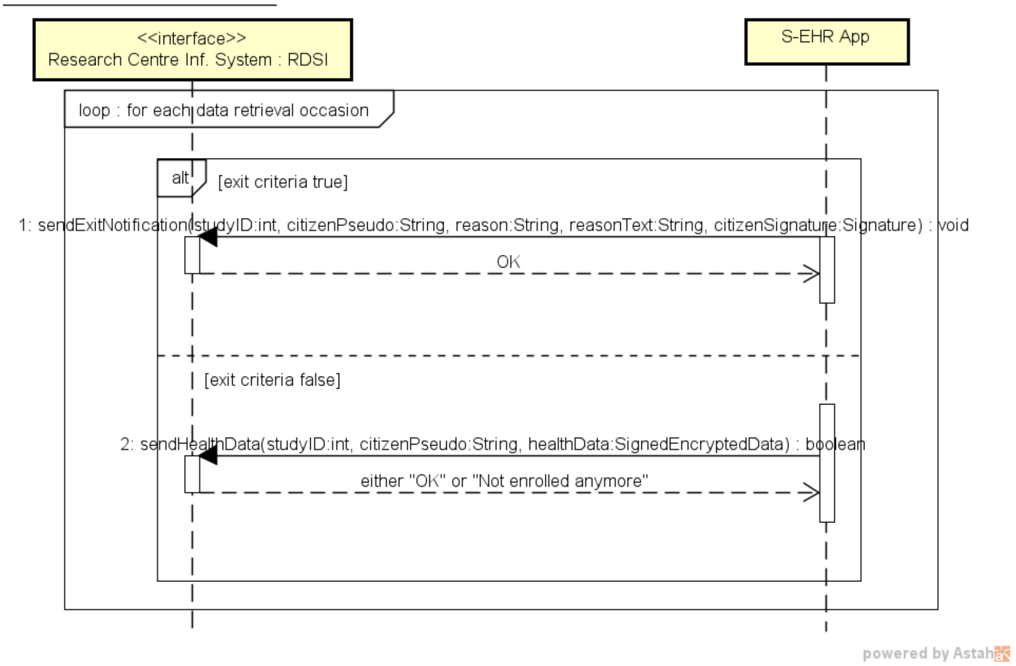

<<interface>>
Research Centre Inf. System : RDSI
S-EHR App
loop : for each data retrieval occasion
alt
[exit criteria true]
1: sendExitNotification(studyID:int, citizenPseudo:String, reason:String, reasonText:String, citizenSignature:Signature) : void
OK
[exit criteria false]
2: sendHealthData(studyID:int, citizenPseudo:String, healthData:SignedEncryptedData) : boolean
either "OK" or "Not enrolled anymore"
powered by Astah

Figure 9 - Sequence diagram for the DATA RETRIEVAL phase

7.5 WITHDRAWAL phase

Purpose: In the WITHDRAWAL phase, the Citizen decides to end his/her ongoing participation in a given study. Upon this request, all further data retrieval operations are suspended, previously collected data are deleted, and the Citizen is deleted from the list of enrolled patients at the RRC.

Actors and components: Citizen, S-EHR App, Reference RC.

Preconditions: The Citizen has been enrolled in the study. At the time of the withdrawal, the study is either still in the enrolment phase or it is already running. It is not possible to withdraw from a study once the data retrieval period has ended.

Steps:

1. The Citizen decides to withdraw from a specific study. In order to do so, (s)he retrieves from the S-EHR App the list of studies to which (s)he is enrolled, and selects "withdraw". (S)he digitally signs the withdrawal.
2. The S-EHR App sends the signed withdrawal message to the RRC.
3. The RRC receives the notification. It acknowledges it, deletes all data retrieved from the Citizen so far, and updates its local list of citizens enrolled into the study.

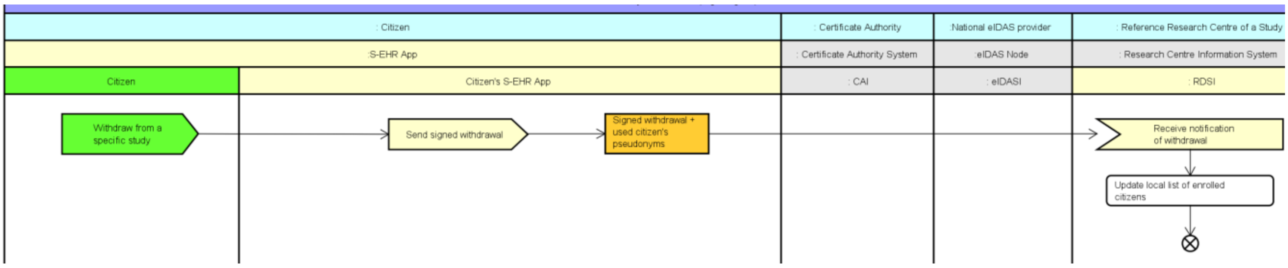

: Citizen
: Certificate Authority
:National eIDAS provider
: Reference Research Centre of a Study
:S-EHR App
: Certificate Authority System
:eIDAS Node
: Research Centre Information System
Citizen
Citizen's S-EHR App
: CAI
: eIDASI
: RDSI
Withdraw from a specific study
Send signed withdrawal
Signed withdrawal + used citizen's pseudonyms
Receive notification of withdrawal
Update local list of enrolled citizens

Figure 10 - High-level data flow of the WITHDRAWAL phase

DRAFT

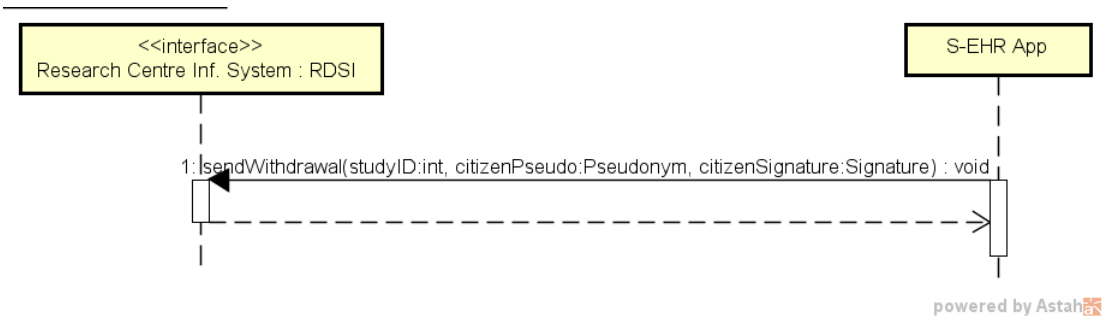

<<interface>>
Research Centre Inf. System : RDSI
S-EHR App
1: sendWithdrawal(studyID:int, citizenPseudo:Pseudonym, citizenSignature:Signature) : void
powered by Astah

Figure 11 - Sequence diagram for the WITHDRAWAL phase

7.6 OPT-OUT phase

Purpose: In the OPT-OUT phase, the Citizen decides to opt out from any future study.

Actors and components: Citizen, S-EHR App.

Preconditions: The Citizen has previously opted in to research studies.

Steps:

1. The Citizen decides to opt out from all future studies. (S)he lets this be known to the S-EHR App.
2. The S-EHR App may consider this decision only for future studies, or also for ongoing studies, in which case a separate withdrawal is necessary for each study. In case there are ongoing studies where the Citizen is enrolled, the S-EHR App prompts the Citizen whether (s)he really wants to withdraw from these ongoing studies as well.
3. The decision to opt out from future studies is stored locally in the S-EHR App. From this moment on, the S-EHR App will not poll the Central Node for future studies anymore.
4. In case there are ongoing studies where the Citizen is enrolled, and the Citizen has explicitly signalled to want to withdraw from these as well, the S-EHR App executes a withdrawal process separately for each study.

DRAFT

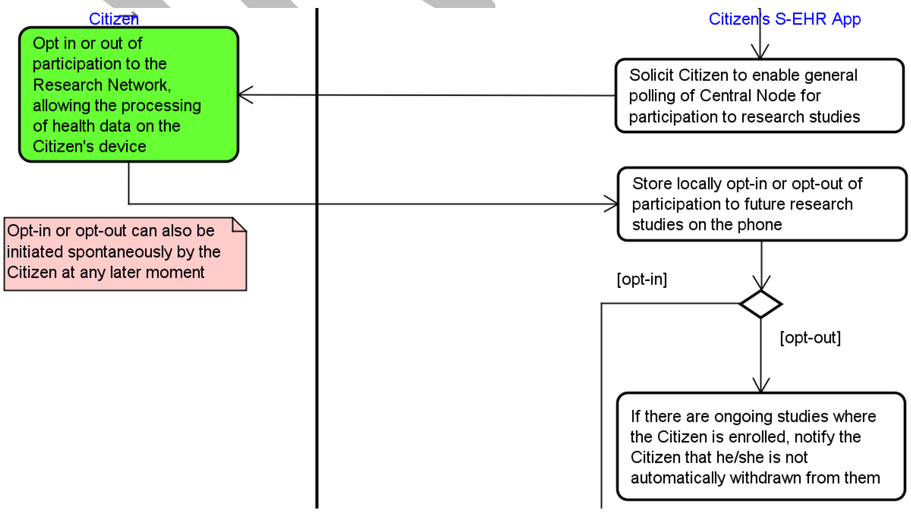

Citizen
Citizen's S-EHR App
Opt in or out of participation to the Research Network, allowing the processing of health data on the Citizen's device
Solicit Citizen to enable general polling of Central Node for participation to research studies
Opt-in or opt-out can also be initiated spontaneously by the Citizen at any later moment
Store locally opt-in or opt-out of participation to future research studies on the phone
[opt-in]
[opt-out]
If there are ongoing studies where the Citizen is enrolled, notify the Citizen that he/she is not automatically withdrawn from them

Figure 12 - High-level data flow of the OPT-OUT phase

8 RELATED WORK

The Protocol described in this deliverable is novel and unique in its approach of retrieving health data directly from citizens' personal devices. The conventional approach so far has been, for retrospective studies, to transform and reuse data from existing sources under centralised control (of a hospital, a region, or an entire country), such as clinical patient data or death records, based on general prior consent given by patients. This is the process assumed and implemented, for instance, in the project *Healthcare Data Safe Havens*, as presented in [HDSH]. In prospective studies, typically a much smaller number (e.g. hundreds) of patients are involved, with consent and data collection happening physically at and carried out by the research centre (or multiple research centres in the case of multicentric studies). The Research Data Sharing Protocol covers both the retrospective and the prospective use cases, but entirely decentralises the data collection process.

In terms of cross-border interoperability, the Protocol adopts the state-of-the-art approach of relying on international data representation standards. Existing projects such as [EMIF], [EHDEN], or Healthcare Data Safe Havens [HDSH] rely on the OMOP CDM standard [OMOP], which is a structured data representation specifically developed for research data interoperability. Our Protocol, on the other hand, relies on the FHIR standard for representing data for research. This choice is not justified by FHIR being inherently better for this purpose—we consider both FHIR and OMOP adequate for the majority of use cases—but by the seamless interoperability it provides with FHIR-based electronic health records, such as the smartphone-based *smart health data* [D2.8]. This way, data can be directly retrieved from an already cross-border interoperable representation, without requiring any complex data transformation that would be tedious to implement on smartphones. However, the rest of the Protocol is designed so that it does not rely on any specific underlying data format: it is up to the implementation to ensure that the query format used in the RDD matches the health record data format on the smartphone, or else to implement data conversion mechanisms.

DRAFT

9 CONCLUSIONS AND NEXT STEPS

This deliverable specifies the second and last major version of the Research Data Sharing Protocol. While this second version introduces several additions and changes, the protocol scope, goals, actors, and its general design have remained the same. The added details and improvements were partly inspired by software design and implementation experience (see deliverables [D4.10] and [D4.17], respectively) and are partly responding to additional requirements formulated by medical researchers.

One element still missing from this deliverable is the specification of the retrieval process of large, unstructured, anonymized documents (e.g. images) by the research centre from the Citizen's hospital.

This requirement and the corresponding solution have been added to the Protocol during the last steps of writing this deliverable. The detailed specifications of the corresponding *R2R Access* protocol (which is a near-identical variant of the *R2D Access* protocol already described in [D4.3]) will be provided as an addendum to this deliverable. This optional extension of the RDS protocol is not required by the InteropEHRate pilots, but could be required, due to current limitations of smart devices commonly adopted by citizens, in more complex research studies. It is expected not to be required in the long term, thanks to the increase of storage and processing capabilities of smart devices.

Similarly to the other communication protocols, the RDS protocol is intended to be supported by different systems, potentially provided by different vendors. Different implementations will be interoperable if compliant to this specification.

For the readers interested in experimenting with the RDS protocol without implementing it themselves, a reference implementation is provided by deliverable [D4.17], including libraries for the mobile device, the research centre, and the Central Node. The design of this specific implementation is documented by deliverable [D4.10].

DRAFT

10 REFERENCES

- **[D2.3]** InteropEHRate Consortium, *Deliverable D2.3—Requirements Specification V3*, 2021. www.interopehrate.eu/resources/*#dels*

- **[D2.6]** InteropEHRate Consortium, *Deliverable D2.6—InteropEHRate Architecture - V3, 2021.* www.interopehrate.eu/resources/*#dels*

- **[D2.9]** InteropEHRate Consortium, *Deliverable D2.9—FHIR profile for EHR interoperability v3*. Please note that the final version (V3) of *FHIR profile for EHR interoperability* is due in December 2021. The previous version of this deliverable, D2.8 - FHIR profile for EHR interoperability - V2, 2020, can be found at www.interopehrate.eu/resources/*#dels*

- **[D3.2]** InteropEHRate Consortium, *Deliverable D3.2—Specification of S-EHR mobile privacy and security conformance levels - V2.* Note that D3.2, final version of privacy and security conformance levels, is due in December 2021. The previous version, D3.1-Specification of S-EHR mobile privacy and security conformance levels - V2, 2020, can be found at *www.interopehrate.eu/resources/#dels*

- **[D3.4]** InteropEHRate Consortium, *Deliverable D3.4—Specification of remote and D2D IDM mechanisms for HRs Interoperability - V2, 2021. www.interopehrate.eu/resources/#dels*

- **[D3.6]** InteropEHRate Consortium, *Deliverable D3.6—Specification of data encryption mechanisms for mobile and web applications - V2, 2021. www.interopehrate.eu/resources/#dels*

- **[D3.10]** InteropEHRate Consortium, *Deliverable D3.10—Design of libraries for HR security and privacy services - V2, 2021.* www.interopehrate.eu/resources/*#dels*

- **[D4.3]** InteropEHRate Consortium, *Deliverable D4.3—Specification for remote and D2D protocols and APIs for HR exchange - v3, 2021.* www.interopehrate.eu/resources/*#dels*

- **[D4.10]** InteropEHRate Consortium, *Deliverable D4.10—Design of library for health data sharing for research v1*, 2021. www.interopehrate.eu/resources/*#dels*

- **[D4.17]** InteropEHRate Consortium, *Deliverable D4.17- Libraries for research health data sharing - V1, 2021.* www.interopehrate.eu/resources/#dels

- **[D6.8]** InteropEHRate Consortium, *Deliverable D6.8—Design of a mobile service for data anonymization and aggregation, 2021. www.interopehrate.eu/resources/#dels*

- **[FHIR]** HL7 FHIR Specifications. https://www.hl7.org/fhir/

- **[Camenisch 2017]** J. Camenisch and A. Lehmann, "Privacy-Preserving User-Auditable Pseudonym Systems," 2017 IEEE European Symposium on Security and Privacy (EuroS&P), Paris, 2017, pp. 269-284, doi: 10.1109/EuroSP.2017.36.

- **[eIDAS 2017]** European Commission — DIGIT Unit D3, eIDAS-Node Installation, Configuration and Integration Manual, Version 1.3, 2017

- **[EJBCA 2021]** PrimeKey, EJBCA WS Support, 2021 https://download.primekey.se/docs/EJBCA-Enterprise/latest/ws/index.html

DRAFT

- **[EMIF]** The *European Medical Information Framework* project, http://www.emif.eu
- **[EHDEN]** The *European Health Data and Evidence Network* project, http://www.ehden.eu
- **[HDSH]** G. Bella et al., *Cross-Border Medical Research using Multi-Layered and Distributed Knowledge.* Prestigious Applications of Intelligent Systems, ECAI 2020.
- **[OMOP]** OHDSI, *The OMOP Common Data Model.* https://www.ohdsi.org/data-standardization/the-common-data-model/
- **[STS1992]** W. Diffie, P. van Oorschot and M. Wiener, "Authentication and Authenticated Key Exchange'', Designs, Codes and Cryptography, 2, 1992, pp.107-125.
- **[1609.2-2016]** "IEEE Standard for Wireless Access in Vehicular Environments--Security Services for Applications and Management Messages," in IEEE Std 1609.2-2016 (Revision of IEEE Std 1609.2-2013), vol., no., pp.1-240, 1 March 2016, doi: 10.1109 / IEEESTD 2016 7426684.
- **[Eckhoff2011]** D. Eckhoff, R. German, C. Sommer, F. Dressler and T. Gansen, "SlotSwap: strong and affordable location privacy in intelligent transportation systems," in IEEE Communications Magazine, vol. 49, no. 11, pp. 126-133, November 2011, doi: 10.1109 / MCOM.2011.6069719
- **[LABIOD2018]** Intercor project, Milestone 5 - Common set of upgraded specifications for PKI and Common Certificate Policy (CP), 2018, https://intercor-project.eu/wp-content/uploads/sites/15/2019/03/InterCor_M5_Upgraded-Specifications-PKI-CP_v1.0_INEA-1.pdf
- **[eHealth2021]** eHealth Network, OUTLINE Interoperability of health certificates Trust framework, v1.0, 202, https://ec.europa.eu/health/sites/default/files/ehealth/docs/trust-framework_interoperability_certificates_en.pdf
- **[eHDSI2021]** eHDSI Business Analyst, Ensure Health Professional (HP) Identification, Authentication and Authorization, 2021, https://ec.europa.eu/cefdigital/wiki/display/EHOPERATIONS/01.+Ensure+Health+Professional+%28HP%29+Identification%2C+Authentication+and+Authorization

DRAFT